Cosimo Distante
Sebastiano Battiato
Andrea Cavallaro (Eds.)

LNCS 8811

# Video Analytics for Audience Measurement

First International Workshop, VAAM 2014
Stockholm, Sweden, August 24, 2014
Revised Selected Papers

# Lecture Notes in Computer Science 8811

*Commenced Publication in 1973*
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

More information about this series at http://www.springer.com/series/7412

Cosimo Distante · Sebastiano Battiato
Andrea Cavallaro (Eds.)

# Video Analytics for Audience Measurement

First International Workshop, VAAM 2014
Stockholm, Sweden, August 24, 2014
Revised Selected Papers

Springer

*Editors*
Cosimo Distante
National Research Council of Italy CNR
Arnesano, Lecce
Italy

Andrea Cavallaro
Queen Mary University of London
London
UK

Sebastiano Battiato
Department of Mathematics
and Computer Science
University of Catania
Catania, Catania
Italy

ISSN 0302-9743
ISBN 978-3-319-12810-8
DOI 10.1007/978-3-319-12811-5

ISSN 1611-3349 (electronic)
ISBN 978-3-319-12811-5 (eBook)

Library of Congress Control Number: 2014953273

LNCS Sublibrary: SL6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Springer Cham Heidelberg New York Dordrecht London



Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

The retail and advertisement industries are becoming more pervasive, with the need for measuring engagement of viewers/shoppers with newly launched campaigns. The Digital Signage sector represents today the third advertising medium in terms of annual revenues after mobile and online advertising. The trend is exponentially increasing and brands, network aggregators, and media planners' needs are moving toward understanding the level of engagement of viewers in order to measure their reaction to new products. While online advertising is mature and has established measurements tools, there are sectors of the sale industry where grabbing anonymous information from the human being is important to measure the effectiveness of a campaign and to take prompt actions to maximize the attention of people to the ad or to the product space. Point of sale and on-shelf solutions are also getting more pervasive due to the needs of measuring how shoppers engage, where attention and gaze estimation in free environment is difficult to perform. Also, measuring the customer experience will allow to stimulate a multidisciplinary approach, which will bring ethologist, psycologist, marketing, and media planner professionals to eventually propose new metrics to study and understand social behaviors of social media.

Video analytics may help in understanding the effectiveness of the branded message by studying and measuring public opinion and polling, geographical concentration of conversation of viewers. To this aim, computer vision and pattern recognition technologies will play an important improvement in audience measurement for its capability to understand several visual cues such as demographics, free gaze estimation, dwell time, emotion and group people proxemics, where low spatial resolution of acquired subjects, changing of the pose, occlusion, illumination changes, large variability of intra-class female age, and ethnicity cohorts represent some critical aspects for recognition.

The aim of this book is to provide an overview of state-of-the-art methods for audience measurements in the retail and Digital Signage sectors, attract end-users, and stimulate the creation of appropriate benchmark datasets to be used as reference tool for the development of novel audience measurement algorithms.

The book is organized into three parts. The first part is an introductory chapter and explains the usage of data for the decision making from a marketing research point of view. It also cover the analytics necessary to the media and marketing professionals to plan, execute, and control their marketing and media actions. This information is useful for computer vision and pattern recognition professionals in order to focus their studies on important data to be extracted from one or more video streams.

Part 2 is about demographics and discusses current and future trends on biometric features used to grub anonymous information about several cues. Several contributions are related to the latest powerful feature extraction and selection methods for gender, age, and ethnicity. Some of them are used both for reidentification purposes for returning viewers, recognition, and to trigger dedicated advertising campaigns to the

current viewer. The contribution of face alignment activity is also investigated and robustness in the recognition process measured. In this context, reidentification is a hot topic since it allows better viewer statistics by filtering repetitions. Besides biometric information grasped from the detected face bounding box, information regarding clothing attribute is also addressed. A contribution related to the use of an RGB-D sensor is also introduced in order to understand which product the viewer is focusing its attention and for how long.

Part 3 is related to modeling consumer behavior by using machine learning techniques. The majority of papers are related to aggregate and analyze demographics information, 3D viewer patterns in order to predict consumer behavior in a retail environment, especially oriented toward the purchase decision process and the roles in purchasing situations.

September 2014

Cosimo Distante
Sebastiano Battiato
Andrea Cavallaro

# Organization

## Committees

### Program Chairs

| | |
|---|---|
| Cosimo Distante | National Research Council of Italy CNR, Italy |
| Sebastiano Battiato | University of Catania, Italy |
| Andrea Cavallaro | Queen Mary University of London, UK |

### Program Committee

| | |
|---|---|
| Dario Cazzato | National Institute of Optics – CNR, Italy |
| Modesto Castrillón-Santana | University of Las Palmas, Spain |
| Oscar Deniz | Universidad de Castilla-La Mancha, Spain |
| Abdenour Hadid | University of Oulu, Finland |
| Robert Ravnik | University of Ljubljana, Slovenia |
| Paolo Spagnolo | National Institute of Optics – CNR, Italy |
| Giovanni Puglisi | University of Catania, Italy |
| Marco Del Coco | National Institute of Optics – CNR, Italy |
| Kenneth Alberto Funes Mora | Idiap Research Institute, Switzerland |
| Dit-Yan Yeung | Hong Kong University of Science and Technology, Hong Kong |
| Marco Leo | National Institute of Optics – CNR, Italy |
| Luiz M. Garcia Gonçalves | Universidade Rio Grande do Norte, Brazil |
| Anil Anthony Bharath | Imperial College of London, UK |
| Giovanni Maria Farinella | University of Catania, Italy |
| Hu Han | Michigan State University, USA |
| Guodong Guo | West Virginia University, USA |
| Pier Luigi Mazzeo | National Institute of Optics – CNR, Italy |

# Contents

# Introduction

# The Applications of Video Analytics in Media Planning, Trade and Shopper Marketing

Matteo Testori(✉)

Dialogica Srl, Milan, Italy
matteo.testori@dialogica.it

## 1 Introduction

The paper intends to highlight the applications of video analytics, particularly the face detection/audience measurement systems in two main areas:

- Digital out of home networks.
- In store analysis – Shopper behaviour and effectiveness of the in-store communications.

The approach is based on field works, cases, and will explain the usage of data for the decision making. It will also cover the analytics necessary to the media and marketing professionals to plan, execute and control their marketing and media actions.

## 2 The State of the Art

Automatic face detection/audience measurement system has been available since 2006/2007. The software is based on algorithm that allows the detection of face/glance, the estimation of gender and age groups.

The technology, based on face detection software, allows to:

- Count the passers-by in a given area (shopping malls, transportation hubs, stores…).
- Measure the dwell time.
- Measure the number of viewers.
- Measure the attention time.
- Split the data by gender and age groups.


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811 pp. 3–20, 2014.
DOI: 10.1007/978-3-319-12811-5_1

The software intercepts the passers-by and, detecting the face/glance, defines the gender and the age groups. The advantages:

Real time counting: like the internet-based analytics, the behavior of the users is measured in real time. The machine based technology allows, at a reasonable price, the 24 h, 7 days detection.

No human interface: traditional marketing research is conducted by interviewers: trained people who meet/interview and observe the customers. The methodology involves several problems, in addition to the well-known opportunities (Deep understanding of characteristics, behavior, motivations of the consumers):

(a) Small samples. A sample of the population/target is selected. Face to face interviews (CAPI – Computer aided personal interviews), Telephone calls (CATI – Computer aided telephone interviews) Web surveys (CAWI – Computer aided web interviews) are conducted and the results are expanded, using statistical inferences models, to the population. The size of the sample, the period, the duration of the interview may affect the deliverables. Small sample means, generally, significant statistical error. Big samples are preferable, but may involve big investments. Of course, the total spending depends also on the methodology that must be carefully chosen depending on the final objective of the research/survey.
(b) Personalism – All traditional methodologies involve the use of a questionnaire and a "relationship" between the interviewer and the customer. The result may be influenced by the quality and the structure of the questionnaire and by the attitude, the capacity and skills of the interviewer.
(c) People may lie: all interviews are based on declaration. For many different reasons people can, and sometimes they do, lie.

Real time detection: traditional surveys are, generally, and mainly in the case of in store/shopper interviews, limited to a short period of time: couple of weeks. This is due to the fact that, in most cases, interviewers cannot stay for long time in a location (Store, Supermarket, Transportation hub…) because of costs and interference with customers. The chance to have some sensors (webcams) installed perpetually in a spot, allows the continuous counting/detection of the consumers. This is extremely effective in the measurement of stochastic/variable phenomenon, like for example, effectiveness of promotions, display, planograms….

From the pioneering commercial software, a lot of new players launched their applications. Most of them are poor in terms of marketing and fails in applications. The main reasons are:

1. No calculation (or very poor/weak/inaccurate) of the traffic.
2. Long detection time.
3. Poor (with high % of error) detection of age groups.

Some other techniques tend to over perform: precision is sometimes the enemy of effectiveness. In statistical terms, it is more likely to have a big and reliable sample instead of having small cases in a very short time. The analysis (we are considering the shopper and in-store research) must capture the real behaviour of the shoppers and doesn't have to interfere with their shopping journey. The shopper must behave

naturally. In this sense, some techniques may be invasive: the eye tracker systems, in fact, involve the usage of wearable hardware on behalf of the shoppers. This implies:

1. Small sample/cases.
2. Small number of outlets.
3. Short period of analysis.
4. Unnatural behaviour.

Eye tracking can be used in laboratory (some big multinational corporations are using it in "fake stores": a reproduction of a supermarket used for testing products, package, displays, and racks), but it is very complicated in real life, although it may add interesting and precise outcomes.

The audience detection systems are also used to track the traffic flown in hubs, outlets, malls, supermarkets. In this sense are more usable, flexible, and cheaper than the systems for fingerprinting the smartphones of the shoppers. They don't require a complicate infrastructure, can be easily installed, and are cheaper. And, most of all, are more sensible and accurate. They can detect the movements in a narrow cone and, in the meantime, can detect gender and age. The wi-fi fingerprints are quite inaccurate. They can track the presence of a smartphone in a wide area, with a precision of some meters. They can track the repeating visits in a shop, can give an idea of the wide movements in big areas, but don't say anything about the narrow location of a shopper, for example in front of different displays in the same aisle. They are useless, for example, in store test when new products must be surveyed.

They have another inconvenience: they are tracking a sample not representative of the whole universe: for example, in Italy, the penetration of the smartphones on the total population is roughly about 50 %. The users are young/adults, medium/high class and high education. The sample is not representative. The risk is to track a portion of the potential target, mostly those who are technology addicted.

The main disadvantage of the face detection systems is that they simply calculate quantitative numbers, they are still poor in age recognition (but, it is quite difficult to discover the difference between two young women, one 17 years old and the other 21 years old). In general, and apart from specific applications, the data from the audience measurement tools, are raw data, with a specific usage in marketing studies.

The data catch the phenomenon, but have limited level of understanding. The meanings behind data are quite obscure, considering that marketing and advertising professionals, in the era of big data, are, more than ever, looking for insights, not for data. The booming of data, mainly from the web and the social media, produced the fever of meanings, insights, and the frightening for tons of meaningless data. Therefore, it is necessary to get raw data from the audience measurement software and derive insight from them.

Another problem consists of the current vocabulary of marketing and media: standard metrics have been used for decades to measure and weigh the effectiveness of media planning and in-store actions. That's the common language to be adopted. Particularly, some metrics are the state of the art:

GRP: Gross Rating Point. It is the index measuring the gross advertisement pressure. It is useful to understand how much pressure is produced on a certain target. It is

expressed in absolute value. So, 100 GRPs are more than 50 GRPs. The GRP derives from the product of the media coverage and the frequency. It shows how many persons are reached by adv (coverage) and how many times they are reached.

This indicator can be calculated in two different ways:[1]

$$Netcoverage\ (^{*}100)\ x\ Average\ Frequency$$
$$Gross\ Impressions/Target\ group$$

Gross Impressions: it derives from the sum of all the audiences generated by the total outputs of a certain advertisement, that is to say, the amount of the contacts developed by one or more adv launched on one vehicle or on a combination of vehicles. In this case, the impressions are qualified as "gross" because this indicator does not take into account the repetition of contact on the same person. The gross impressions are expressed in absolute values.

Net Impressions: it indicates the number of every single person exposed at least once to the advertisement.

The net impressions are calculated with the following formula:

$$Gross\ impressions\ /\ Frequency\ =\ Net\ impressions$$

Reach: the number of different persons reached by an advertising plan and once exposed to the advertising.

Average Frequency: it indicates how many times each person belonging to the target is exposed on average to the advertising.

Media coverage: the average number of persons (expressed in %), belonging to the target group, actually reached by advertising.

Some other topics have to be taken in consideration: we are facing a lot of noise on web analytics. Big Data are promising to unveil all about the web-aholics. This is fine. In 2013, 10% of the total USA retail (Forrester research) turnover came from the online, with a growth of 10% versus 2012.

E-commerce is growing. Off line retail is slowing down. This is a fact. But, today, most companies survive thanks to the traditional/off line commerce.

A paradox: we know everything about 10% of the total retailing expenses (Growing, future, booming…) but what about the everyday business? The Amazon street is not the main street, yet.

Web analytics are the epitome of massive, intensive, clear, accurate data for targeting and understanding the behaviour of the consumers and shoppers. No other media, as from the common sentiment, has such a rich and exhaustive set of data to understand the target, almost in real time.

Any metric on new media (and the digital out of home is, in fact, a new media) cannot take in consideration that, at least at a superficial glance, will be compared to the web analytics.

[1] Net coverage = Net impressions/Target.

Another fact, much more significant than the fashionable mood regarding web data, is about the comparison among different media. Any analysis on effectiveness and efficacy of a media mix must consider that any single media is compared to the other ones, in order to have a feedback/result on its effectiveness.

Some findings:

In the UK, 20 million people visit every week Tesco, the first retailer in the United Kingdom. They are the double of the audience of X Factor. In the United States 80 million people watch the Super Bowl (the final championship of the National Football League American Football) one night a year. 140 million people enter every week Walmart stores (the first retailer in America and in the world). In 2007, Tesco had 1,500 supermarkets in the UK. Today 3,054. The turnover of Tesco is equal to the GDP of Vietnam. The turnover of Walmart is equal to the GDP of Sweden. It is estimated that in 2010 Procter & Gamble spent U.S. $3.5 billion in trade and shopper marketing in the United States. The total size of the investment in shopper marketing in the U.S. is estimated at $35 billion per year, and is still growing. From 2004 to 2010 the market grew, in the U.S., an average of 21 % each year. More than social media.

On average, the time spent in front of a shelf is approximately 15 s. On average, one shopper, in fast moving consumer goods, looks at the shelf for about 4 s during which chooses what to buy.

This findings prove that a store, a shopping mall, a supermarket is a media. Extremely important, if we consider that any advertising inside the store is aired where the shopper shops. The ad is close to the purchase. According to POPAI (International Association that aims to promote the culture of point of sales), more than 70 % of purchase decisions are taken inside the store. Other companies and institutions have presented different percentages (usually higher), but the data of POPAI is the most well known and reported in articles and conferences. The complexity of the purchase process must be detected in order to better understand the path from the desire to the purchase. A simple (and very simplified) process, called path to purchase, is commonly used in literature and in the marketing practice, to follow the funnel from intention to purchase:

1. Awareness
2. Engagement
3. Discovery
4. Investigation
5. Selection
6. Purchase

The funnel paradigm of purchasing behavior, which began with awareness and ended in purchase, was an early shopper marketing model. It helped brands think about the shopping process in a more refined way. The shopper marketing was defined as "*The use of strategic insights into the shopper mindset to drive effective marketing and merchandising activity in a specific store environment*" [17].

But the paradigm was disrupted by digital technology. Shopping, investigating, considering, selecting and buying can be done anywhere at any time. It is possible to look for one brand online, and then become aware of another instantly via pop-up

advertising, derailing the process. Or be offered a coupon by a friend on Facebook. Currently the decision journey can be described with the following steps:

1. The consumer starts considering a set of brands basing on brands awareness, perception, and exposure to a multiple (and entangled) set of touch points: the traditional media, the web, the word of mouth (physical and virtual/social), web surfing, the stores (on/off line)…
2. The information journey, on and off line, helps the consumers/shoppers to add or subtract brands, as they evaluate what they want and what they can afford.
3. The shopper selects the channel: on line or off line. Surfing the net helps to get information on the product features, the pricing, the evaluation from other shoppers.
4. The shopper selects the store: the result depends on several factors, all peculiar to the kind of product, the past experience, the expectations, the perceived value, the pricing, the service.
5. The shopper buys the product
6. The shopper tries the product: that's the stage of real physical experience and the first step for repeated purchase.
7. The shopper experiences the after-sale service.

Apart from the funnel/steps that can be interpreted in several ways, the overlapping of digital and physical shopping journey, forces scholars, marketers, researchers, to find a common ground to evaluate the metrics to measure the path to purchase. The face detection systems can be used as a basic set of metrics to develop a comparable set of indicators to map the decision journey that involves different touch points.

Taking in consideration the media and the stores, some basic topics have to be considered:

- A brand must be noticed/seen: no matter if we talk about traditional adv, or web adv, or in-store visibility. In the case a product/brand is not able to attract the attention of the shopper, there is no chance to pass from the attraction to the purchase. The first step in the funnel is the ability of the brand to catch the attention. It is the initial condition to create awareness, desire, engagement, purchase.
- Attraction is not enough: a brand must be able to keep the attention, after an initial sight. Attention is the marker of the interest. The more attention a shopper pays to a brand, the more interest grows.
- The initial attraction and attention must turn into interest: if a shopper is not interested in a brand, it is almost impossible that an initial attention becomes real attitude to purchase. A brand must be relevant to the shopper, to turn a superficial attention into an act of purchase.
- In addition, a brand, to become a real "like" product, must engage the shopper.

All the above attributes (and many more like, for example, the brand awareness, the brand equity, the brand desirability) must be measured to track the path to purchase in a multi-channel/media market.

## 3 Sum-Up

1. Video analytics can be a powerful tool for marketing research.
2. The system are easy to use and cheap.
3. They are reliable (but must be tested before the usage to guarantee the quality of the measurement).
4. They can provide real time and 24 h data.

But, to be really usable by marketers and advertisers, they must:

- Be used as raw data to feed an exhaustive and understandable set of analytics.
- The analytics must be similar and comparable to the standard metrics currently used by marketing and advertising.
- The analytics must make different media and locations fully comparable, by having common indexes.

## 4 The Analytics Based on Face Detection Raw Data

Based on the above considerations, and referring to the standard metrics/ratios used in advertising and marketing, it is possible to get some ratios to describe the behavior of the shoppers, the effectiveness of advertising, shelf space, product displays, visual merchandising. Some of them are the basis in the analysis of all media, no matter if we consider a store, a display, a digital out of home network. A brief description follows:

Attraction Index - it measures, in %, the number of viewers, that is to say, passers-by looking at a POC (point of contact). For example, "45 % viewers" means that 45 % of passers-by looked at the POC.

Attention Index - it expresses, in %, the amount of time (seconds) a passer-by devotes to the observation of a POC (Point of Contact).

This index is generally used in combination with the "Attraction Index" in order to evaluate the audience level of interest.

Relevance Index - it indicates how much a POC is perceived as interesting. The more this index rises, the more the interest level turns up. For instance, a Relevance Index equal to "0.8", shows a level of interest that is twice the index equal to "0.4".

Engagement - it indicates the level of engagement to a POC. It derives from the correlation between the Relevance Index and the Attention Index, and is based upon the axiom, for which a certain piece of advertising generates interest if it attracts and the viewer devotes to it a certain amount of time. It is expressed by a range of values from −1 to +1. The more this value approximates to +1, the more the POC generates interest/engagement. If the value approaches to "0", it means that the POC is neuter. The more the value approaches to −1, the more the POC doesn't generate interest.

More specific indexes are used in supermarkets and store outlets to track the path to purchase:
Conversion index - it measures (%) the number of entrances in a given point of sale on the total number of passers-by outside. For example, an index equivalent to 50 % means that one passer-by on two went in actually.

Sale index - this index is used for different purposes, for example with the aim to measure the commercial result, by showing the kind of impact produced by the traffic flow on the number of purchase acts.

Store potential Index - it is a heterogeneous index that expresses the potential of every single shop in a retail chain, as an effect of the location, the attraction index, the conversion index, and the purchase acts.

Traffic Index - in reference to a shop or a retail chain, it expresses the effectiveness of the location in relation to the outside traffic flow. This index is useful to evaluate the impact produced by the location of the point of sale, that is to say, how the location of the shop contributes to the entrance rate (conversion index) and the sales.

Some more are very specific to the Digital out of home:

DRP (Digital Rating Point) - it is the indicator of advertising effectiveness (developed by Dialogica) for the measurement of Digital out of Home. It measures the effective reach and exposure. It is expressed in absolute values, and it is much inferior to the GRP, as a consequence of the fact that this indicator considers just the real exposure (the actual viewers), not the gross or net impressions.

ATE - ("Actual Time of Exposure"): it measures the actual reach and average time a viewer is exposed to the message.

The usage of the ratios, in addition to the sell-out data in case of retail outlets and supermarkets, allows a comprehensive understanding of the in-store path to purchase, the effectiveness of advertising and the comparison of the Digital out of Home networks with other media.

## 5 The Digital Out of Home

If we consider the digital communication, again, we can count on tons of data (Google analytics, web research) on the web but new media, the Digital out of home (video out in the streets, in shopping malls, in transportation hubs…) suffers because of the lack of data and insight.

"*The medium is practically invisible with media strategists and planners, the key decision makers who control advertising budgets. It's not that they don't know it exists. The medium is simply too difficult to plan and buy. Media Planners and strategists have had difficulty recommending digital place-based media to their clients for many reasons, including structural issues and lack of data*" [18].

In USA, more than 450 Digital out of home networks (Dooh) are active, with more than 1.000.000 screens. Most of the networks survive thanks to the advertising

investment from brands. Advertisers used to plan their investments on data, analysis, ratios, very common in broadcast media, like Television, Radio, and Web.

A paradox: the development of the Web has produced an excess of attention and immense data: the investments have diverted resources to the media and traditional sales channels that, despite the growth of the web, are still the most important marketing vehicle for companies.

The DooH (Digital out of home) is considered as a technological evolution of the traditional out of home advertising. The paper posters will be transformed (at least partially) in digital screens, where contents (Static, dynamic, interactive, video, audio, web contents) are managed from a remote central content management software. Advertisers tend to consider DooH as a branch of the traditional out of home media. This is changing because of the availability of data: the usage and the elaboration of audience detection data with the development of analytics and the possibility to have the ordinary metrics used by advertisers and media planners (GRP, Gross and net impressions, frequency) is revolutionary.

The consequences have tremendous impact on the industry: in the following chart the evolution of media spending (in Italy) is reported (Table 1):

**Table 1.** The media investment in Italy – 2010/2013

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TV | 49.6 | 49.2 | 52.8 | 52.4 | 53.5 | 54.0 | 52.5 | 51.0 | 51.5 | 52.1 | 53.4 | 53.5 | 52.2 | 52.6 |
| Web | 1.3 | 1.3 | 1.2 | 2.2 | 2.4 | 2.7 | 3.8 | 5.4 | 6.4 | 9.6 | 10.8 | 12.6 | 15.7 | 18.9 |
| OOH | 8.9 | 9.0 | 8.4 | 8.3 | 8.1 | 8.0 | 7.9 | 7.7 | 7.7 | 7.0 | 6.7 | 6.1 | 5.9 | 5.7 |

Source: Group M

The out of home market lost more than 3 points from 2000: currently TV and Web counts for more than 70% of the total advertising spending. The opportunity offered to the DooH network, thanks to the new analytics, is to get out of a declining market (out of home) and grab investments from other media: TV but, primarily, the internet. All analytics developed starting from Audience measurement data give advertisers and media planners the opportunity to treat and evaluate DooH exactly like the web in their plans.

We underline, again, the fact that raw data from audience measurement tools are not enough. They don't give planners the metrics they need.

An additional revolution in media is affecting the industry: programmatic buying is a new wave in media planning. Web advertising space is managed automatically; some software/platforms surf the inventory of web adv spaces to optimize the media planning/effectiveness and investment. The available inventory may be auctioned depending on available (or not yet sold) space.

Big corporations, like Procter & Gamble (The biggest advertising investor worldwide), have developed their own programmatic buying platform. They recently reported that their goal is to manage 70% of their total digital investment programmatically. Also big media agencies, like WWP and Publicis, are moving the same way.

In this scenario, data and analytics will drive the dance. Media planners must optimize the investment and effectiveness on a given target. A typical goal for a planner is, for example: optimize advertising pressure (measured by the GRP's) on men, age 18/54, with high income and status.

Therefore, it is necessary to merge the info on status and income (and get the standard age brackets) with data collected through audience measurement system (Table 2).

**Table 2.** Standard age brackets for media planning

| 14/17 | 18/24 | 25/34 | 35/44 | 45/54 | 55/64 | 65/74 | >74 |
|---|---|---|---|---|---|---|---|

The necessity of the media planners is satisfied by merging the Audience measurement data with existing traditional surveys, made by marketing research institutes. The interviews (CAPI: Computer aided personal interviews. CATI: Computer aided telephone interview. CAWI: Computer aided web interview) collect a wide and deep set of data: socio demographics, income, status, education, behavior, consumptions… and add all information regarding the media/marketing target missing in case of data from Audience measurement. The minus of the interviews stands in the relative small size of the sample, in the instant (not continuous) collection of data. On the other hands, the minus of the interviews is the plus of audience measurement data. The merge of data from different sources solve the problem.

The final consideration: Digital out of Home has the same data of the web and can enter both in traditional media plans and in the new programmatic platforms. This is a real revolution for the industry: in the (very near) future, the lack of reliable analytics will exclude the DooH networks from the arena of advertising investments.

## 6 Case Study

The case reports the experience of the Grandi Stazioni Digital Network in Italy. Grandi Stazioni Media is the company that manages the advertising and the stores in the 14 biggest railways stations in Italy. Since 2013 Grandi Stazioni installed the audience measurement software in more than 130 interactive totems in 5 stations. All data come from the network.[2]

Some basic numbers: raw data from the audience measurement (average month):

Traffic: 16.900.000
Viewers: 5.445.000
Viewers male: 56%
Viewers female: 44%
Child: 3.2%

[2] We thank Grandi Stazioni for the audience measurement data, and GfK Italy for data regarding the interviews conducted in the stations.

Young: 71.8%
Adult: 24.5%
Senior: 0.5%
Average Dwell time: 5.6″
Average Attention Time: 1.8″

Raw data represent a big jump in understanding the traffic, the attention, versus previous evaluations. But don't make the media usable in advertising planning and in programmatic platforms. In addition, no outcomes are available regarding effectiveness, attraction, attention, relevance, engagement. We can add, using raw data from the audience measurement software, additional information:

Attraction: 33%
Attention: 27%
Relevance: 3.61
Engagement: 0.75

The analytics report that 33% of the passers-by look at the screens, they spend 27% of the total time looking at the video. The ads are relevant and engage the passers-by (the maximum level of engagement is 1). But, to make the network usable for media strategists and planners, it is necessary to calculate the net impressions. We derive them starting from the calculation of the frequency (to be calculated for every single hub/station):

Gross impressions: 16.900.000 (same as traffic)
Frequency: 2.8
Net Impressions: 6.035.714

Standard analytics are now available: an additional step involves the target segmentation. It is necessary, to avoid the limits of the audience measurement data, to merge the results with other sources (in the case, taken from GfK interviews) (Fig. 1).

Raw Data | Analytics | Graphs - Index | Correlations | Graphs - Correlations | GFK Statistics

| | Degree | High School | Junior High School | Primary School |
|---|---|---|---|---|
| Education | 56.283,00 | 67.776,00 | 26.213,00 | 0,00 |

| | Freelancer | Artisan | Manager | Employee | Workman | Housewife | Student | Retired | Unemployed |
|---|---|---|---|---|---|---|---|---|---|
| Profession | 17.092,00 | 15.607,00 | 6.425,00 | 49.247,00 | 27.186,00 | 0,00 | 10.925,00 | 0,00 | 23.875,00 |

| | Upper | Upper Middle | Middle | Middle To Lower Middle | Lower Middle | Lower |
|---|---|---|---|---|---|---|
| Family Condition | 6.188,00 | 22.785,00 | 95.646,00 | 70.612,00 | 2.901,00 | 2.512,00 |

| | Upper | Middle | Lower |
|---|---|---|---|
| Status | 28.579,00 | 99.385,00 | 22.459,00 |

| | Lower | Middle Lower | Middle | Upper Middle | Upper |
|---|---|---|---|---|---|
| Income | 9.661,00 | 36.039,00 | 51.448,00 | 33.222,00 | 20.053,00 |

**Fig. 1.** Detailed info (from GfK) on target

We can consider one typical example of media planning; the planner has the goal to direct a campaign to the male target, from 25/44 years, high income. Data can be extracted from the Grandi Stazioni Database and analytics are calculated:[3]

## 7 The Retail Outlets

As we said, a store is also a media. It is the place where the brands meet their clients, the place where all the efforts of marketing and advertising find their final destination. All marketing strategies and actions are focused on the final choice of the shopper, influenced by the awareness of the brand, its desirability, its equity. Most of the final purchase decision is taken inside the store but, again, a funnel leads the shoppers from the simple awareness, through desire, attraction, relevance, engagement, to the purchase. The measurement of the funnel, and the impact of the in-store communication, can be measured with the aid of the audience measurement data. Also in the case of the retail outlets, data must be managed and merged with other sources, typically the sales data. In addition, outrageous amount of money is invested in the visual merchandising, in store advertising, to favor the customer/shopping experience, to attract the shoppers, to communicate the brand. The return of such an investment is quite cloudy. No specific methodology has been developed to calculate the real return of the in store advertising.

Different locations, different merchandisings have significant impact on sales. Some basic questions, in general, have to be answered:

- Is the store in the right location?
- Is the store attractive enough to passers-by?
- Is the in-store communication effective in stimulating and attracting customers?
- What percentage of passers-by enters the store?
- Are people who enter the store within the target?
- How many people look at the products on display?
- How many people make a purchase?
- Is the sales team a good fit for the store?

Thanks to the face detection technology it is possible, if data are reliable, and used in a scientifically correct model, supported by high number of cases and observations, in different environments, answer to the questions. Straight to a case: a network of outlets in different locations. Each outlet was equipped with 3 audience measurement software: 2 installed in the front windows, one inside the store. The systems collect the traffic flown, attention time, dwell time. The results are correlated to the sell- out data. It is possible to define the store potential index in a network, on the basis of its location, attractiveness, entrance flown, and attention to the products.[4] The stores engage young shoppers who enter the stores. The retailer also provided sale data divided by gender:

[3] Analysis, indexes, ratios, analytics from Dianalytics™. Dianalytics is a patented and registered property of Dialogica. Audience data come from Quividi audience measurement tools.

[4] All data are collected using the Quividi systems. The analysis, correlations, regressions, indexes are from Dialogica.

in this case it is possible to compare the traffic outside, the people inside, with the purchase, dividing all data by gender (Fig. 2).

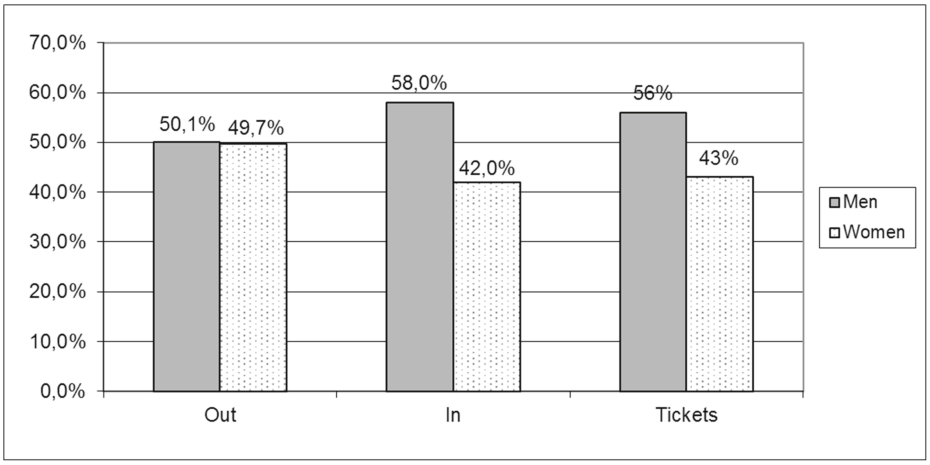



**Fig. 2.** Chart 1: Shopper segmentation – out/in and tickets

The chart 2 compares the trend of the number of viewers outside the stores with the number of people inside and the total sales (Fig. 3). Some further analysis is possible. By calculating the sale index (see p. 8) on passers-by outside the store and the shoppers who actually entered, it is understandable that, on average, 3.5 % of the passers- by (outside) made a purchase, versus 16.6 % of the people who entered the store.

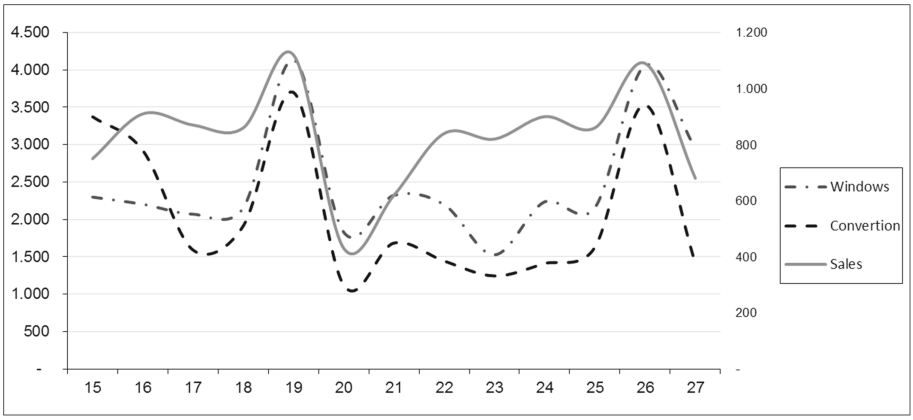



**Fig. 3.** Chart 2: Windows' viewers, conversion ratio, sales

The data are useful to compare different locations, layouts, creativity in the windows and evaluate the effectiveness on conversion and sales, even dividing the outcomes by gender and age (Fig. 4).

Further analysis, using appropriate methodology and systems, may involve correlation and regression among variables.

The chart 4 reports the correlation between sales and the attraction index for men (Fig. 5):

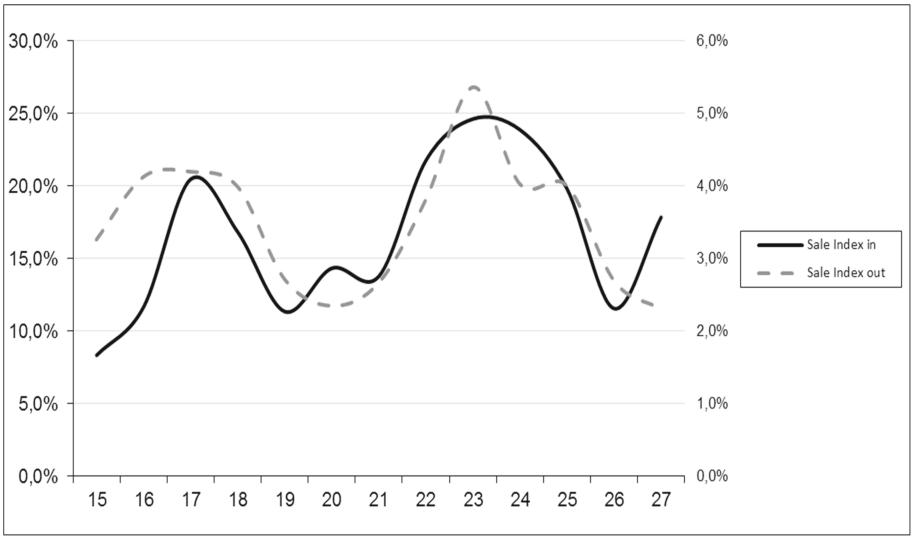



**Fig. 4.** Chart 3: Sale index – traffic outside/inside

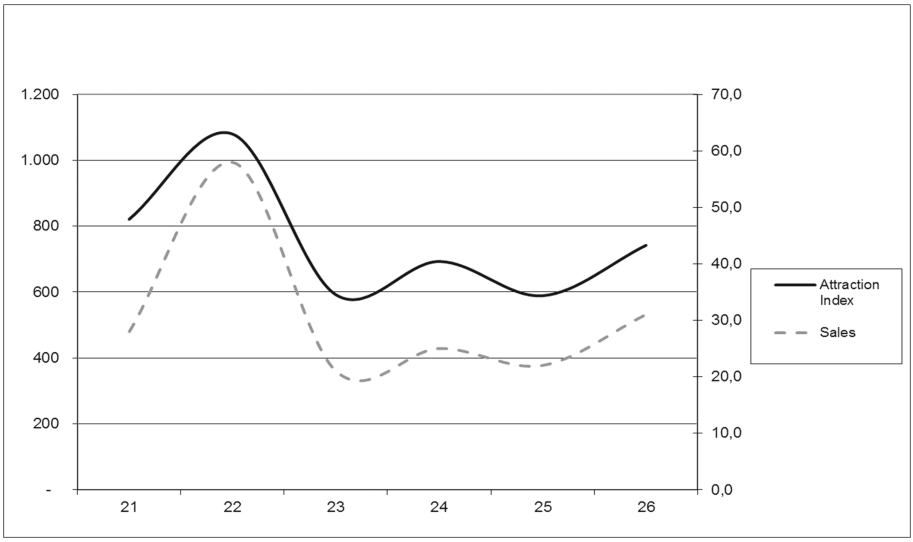



**Fig. 5.** Chart 4: Attraction index and sales – men

The path to purchase, the behavior of the shoppers can be detected using the metrics already met (Table 3):

**Table 3.** Path to purchase and shopper behavior – gender

| | Attraction | Attention | Relevance | Conversion | Sale index | Engagement |
|---|---|---|---|---|---|---|
| Men | 44 % | 26 % | 44 % | 17 % | 3.9 % | 0.73 |
| Women | 46 % | 24 % | 45 % | 15 % | 2.9 % | 0.81 |

In case of Men, the highest factor correlated to the sales (0.98 − $R^2$: 99) is the viewing of the outside windows. Similar results have been recorded for women, with a difference. In case of women the correlation Dwell time/sale is significantly high (0.82 − $R^2$: 70), like the attention time (0.75 − $R^2$: 50). The same metrics collapse for men (Correlation: Dwell/Sales: 0.17 – Attention/Sales: 0.10).

## 8 The Supermarkets

In the case of supermarkets, similar considerations to the retail outlets can be made. But, the complexity is much higher: more traffic, different flows, huge number of products, immense stimuli, and different locations for the same brand inside the store. A topic, in the current depression of consumption and in the growing wave of the e-commerce, is the effectiveness of displays, inventory, layout: practically, the best possible shopping experience and the best possible revenues for retailers and brands. A recent research in the UK[5] (sample: 100 retailers and 500 consumers) reports that 92 % of the retailers think that customer experience is vital for the business, but 26 % has no specific strategy, and 30 % of those who developed some customer programs declares that the results are under their expectations.

31 % of the consumers left a retailer because of a poor customer experience. 56 % of the retailers are not available to identify the improvement of the in-store customer experience. Technology is perceived as a very important tool, but the way to use it to increase the overall satisfaction is still critical. 82 % of the retailers think to offer a positive experience, but 72 % of the shoppers declare a poor, negative service/experience, both in the on-line and the off-line shopping.

Video analytics data, combined with sales data can give the retailers and the brands several information (again, the focus is on information, not on data) to improve the relationship with the shoppers, the customer experience, the revenues.

A classic topic: the effectiveness of the shelf exposure/planogram/inventory. We report a case where a category has been analyzed by installing the technology on the shelves of 4 stores.

[5] Great Expectations – Qmatic, May 27, 2014 - http://qmatic.com/en/Blog/Great-Expectations-a-New-Qmatic-Research-Report/.

Some data:

Total universe (Weekly shoppers of the retailer):1.600.000.
Average traffic detected into the stores (Weekly):18.700.
Standard error: ±1.42%.

Due to the fact that the face detection systems[6] count the passers-by every opening day, the sample is big enough to get relevant and reliable findings, even with a limited number of outlets.

Additional information must be provided by the retailer:

Sales per SKU (Stock keeping unit – Single product/barcode).
Space (cm.) per SKU.
Nr. of facing (nr. of single visible product/brand on the shelf).
Level (Level in the display, from bottom to top).
Price.
Discount/promotions.

All data are compared and analyzed: before proceeding with the analysis it is necessary to make the data comparable by:

1. Defining the common inventory in all the stores (some products can be delisted in one store, but still on the shelf in the others).
2. Defining the baseline of sales (regular sales, net of promotional efforts). In this case a linear regression model has been adopted.
3. Defining the seasonality of sales: in the case the methodology of simple moving average (SMA) has been used.
4. Defining the uplifts, as a difference from actual sales and baseline.

Having aligned the data it is possible to get insights.

The category has been detected for 8 weeks, in the as-is structure (the actual status in terms of inventory, space, price, promotions, level, facing). Some findings:

- 9% of the store traffic reaches the aisle/display.
- Inefficient display: best performing products have not enough space (risk of out-of-stock and poor visibility).
- Private labels (The brand of the retailer) are underperforming because of lack of space/visibility.
- 15% of the inventory is generating poor sales and margins.
- Attraction is quite good: 45% (45% of the passers-by looking at the shelf).
- Attention: 29%.
- Relevance: 45.2.
- Sale index (Total category): 39.8%.

The planogram has been redefined by delisting the underperforming products, listing 4 new SKU (From the market leader), adding space to the best performing

[6] Measured by Quividi, analytics and insights from Dialogica.

products, placing at the best level the most important pack size (the volume generator), and, in general, making all the planograms more clear and simple (horizontal display of products) (Fig. 6).

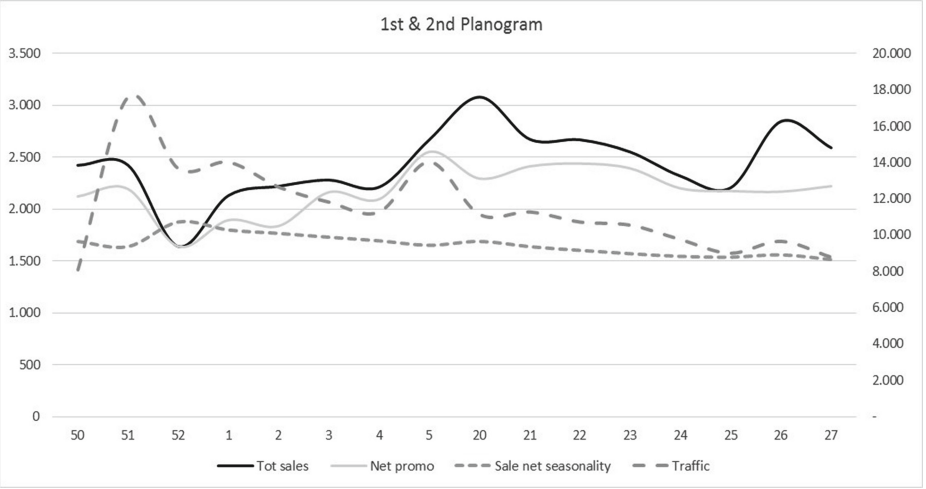



**Fig. 6.** Chart 5: Effectiveness of display – case ½

In spite of a decline in traffic, in general and net of seasonality, the number of viewers is slightly declining and potentially growing (trend). The display is much more effective, with a growth of 10 points in terms of attraction. This is mainly due to the best space allocation and shelf positioning of the leading brands, the market leader and the private label. Sales, in general and net of promos, are growing (Table 4).

**Table 4.** Results

| | 1st Planogram | 2nd Planogram | Δ | Trend |
|---|---|---|---|---|
| Traffic | 103.000 | 80.944 | −22 % | −11 % |
| Viewers | 46.615 | 44.201 | −5 % | +14 % |
| Attraction | 45.7 % | 55 % | +10 | |
| Sell-out (total) | 18.000 pcs. | 20.939 pcs. | +16 % | +31 % |
| Sale per single SKU | 28.9 pcs. | 32.4 pcs. | +12 % | |
| Sale per SKU (net promotions) | 27.7 pcs. | 28.8 pcs. | +4 % | |

## 9 Conclusions

Video analytics/face detection systems can be a very powerful new tool for marketing analysis. The availability of data in real time, the sample size, the possibility to measure in real time the audience, the consumer behavior, offer huge opportunities to find useful

and profitable insights. In the case of the Digital out of Home, they represent a real turnaround in terms of measurement, analytics, and insights and allow the media to be fully considered by planners, advertisers and media agencies.

In case of the store analysis the face detection systems offer new opportunities to test, measure, plan, execute and control the real effectiveness of marketing and advertising actions, calculating the ROI (Return on Investment).

Raw data must be enriched with analytics to get the insights. All the process must be managed professionally, by managers, consultants, researchers, with appropriate knowledge and skills in marketing, advertising, trade, sales and marketing research. Analysis involves the usage of statistical test/models that must be carefully used and findings must be understood and interpreted correctly.

The data are actually matched to the web analytics, making data from different sets (the web, the Dooh, the stores…) comparable.

## References

1. Balconi, M., Antonietti, A.: Scegliere Comprare: dinamiche di acquisto in psicologia e neuroscienzs. Springer, Milano (2009)
2. Corstjens, J., Corstjens, M.: Store Wars. Wiley, Hoboken (1995)
3. Desforges, T., Antony, M.: The Shopper Marketing Revolution. Round Table Press, Highland Park Il (2013)
4. Fornezza, F., Testori, M.: Ecologia della Marca. IPSOA, Milano (2009)
5. Hritzuk, N., Jones, K.: Multiscreen Marketing. Wiley, Hoboken (2014)
6. Lugli, G.: Neuroshopping. Apogeo, Milan (2010)
7. McDonald, C.: Advertising Reach and Frequency: Maximizing Advertising Results Through Effective Frequency. NTC Business books, Lincolnwood (1995)
8. Meroni, V.M.: Marketing della pubblicità. Il Sole XXIV Ore, Milano (1990)
9. Ogilvy, D.: On Advertising. Carlton Books, London (2007)
10. Ogilvy, D.: Confessions of an Advertising Man. Southbank Publishing, London (1963)
11. Ries, A., Trout, J.: Positioning. McGraw-Hill, Milan (1981)
12. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
13. Shannon, C.E., Weaver, W.: La teoria matematica delle comunicazioni. Etas Kompass, Milano (1971)
14. Sissors, J.Z., Baron, R.B.: Advertising Media Planning. Mc Graw Hill, New York (2010)
15. Sorensen, H.: Inside the Mind of the Shopper. Wharton School Publishing, University of Pensylvania, Upper Saddle River (2009)
16. Young, C.E., King, P.D.: The Advertising Research Handbook. Ad Essential, Seattle (2008)
17. The Path to Purchase Institute, p2pi.org, http://www.p2pi.org
18. Digital place-based Media Trend Report – Screenmedia Daily, December 2013. http://www.screenmediadaily.com

# Demographics

# Pervasive Retail Strategy Using a Low-Cost Free Gaze Estimation System

Dario Cazzato(✉), Marco Leo, Paolo Spagnolo, and Cosimo Distante

National Research Council of Italy - Institute of Optics, Arnesano, LE, Italy
{dario.cazzato,marco.leo,paolo.spagnolo,cosimo.distante}@ino.it
http://www.ino.it/en/

**Abstract.** This paper proposes a pervasive retail architecture based on a free human gaze estimation system. The main aim of the paper is to investigate the possibility to automatically understand the behavior of the persons looking at a shop window: this is done by a gaze estimation technique that uses a RGB-D device in order to extract head pose information from which a fast geometric technique then evaluates the focus of attention of the persons in the scene (even more persons at the same time). The main contribution concerns with the application into this challenging research field of a gaze estimation working without any initial calibration and, in spite of this, able to properly deal with completely unaware persons moving in unconstrained environments. Preliminary experiments were conducted in our lab in order to quantitative validate the accuracy of the gaze estimation on different benchmarks of persons. Then the qualitative evaluation of the effectiveness of the proposed architecture was conducted in a real shop window demonstrating the ability to deal with real challenging environmental conditions.

**Keywords:** Free gaze estimation · Human-computer interaction · Focus of attention · Pervasive retail · Digital signage

## 1 Introduction

Gaze tracking plays a fundamental role to understand human attention, feelings and desires [17]. Automatic gaze tracking can open to several application in the fields of human-computer interaction and human behavior analysis, therefore several techniques and methods have been investigated in recent years. When a person is in the field of view of a static camera, gaze can give information about the focus of attention of the subject, allowing for gaze-controlled interfaces for disabled people [22], driver attention monitoring [10], pilot training [33], provision of virtual eye contact in conferences [32] or for the analysis of marketing strategies [25]. Concerning the last, consumers' visual attention estimation can lead to understand underlying display organization, define the optimal disposition of product in shelves, conduct statistics about most interesting products and several other applications. Therefore, several works that make use of computer


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 23–39, 2014.
DOI: 10.1007/978-3-319-12811-5_2

vision and pattern recognition algorithms have been presented in the last few years, as well as design principles to reproduce intelligent environments that are sensitive and responsive to the presence of users and their environment on a large scale, like in [9]. In [26], a camera enhanced digital signage display is presented. The system can extract metrics like person's dwell time, display in-view time and attention time. For the last, a multi-view Active Appearance Model (AAM) registration method is used to estimate head orientation. The work of [31] head pose is used to infer people's visual focus of attention in dynamic meeting scenarios. In [29], a study about searching for a target by consumers performed by using infra-red (IR) markers and an head-mounted eye tracking system is proposed. A review of the usage of commercial eye-tracker in marketing analysis can be found in [25].

Most of the works in the state of the art uses only a face detection algorithm to determine whether observers are facing the object of interest, or a discrete head pose estimation procedure to reveal the macro-area of interest. Moreover, works that try to understand the focus of attention for an environment make use of a commercial eye-tracker, making the overall cost of the system prohibitive for the retail market. Instead, works that use low-cost systems try to understand the focus of attention on a single object, like a target or a screen. This paper presents an intelligent shop window architecture for indoor environments able to understand where persons are looking at. The architecture makes use of an RGB-D device in order to extract head pose information as the input for a fast geometric gaze estimation technique that evaluates the focus of attention of the persons in the scene. The contributions of the work under consideration are that:

- an estimation of the gaze ray for users that are looking on a shop window and understand the observed object is proposed;
- the presented system is low-cost and makes use of a commercial depth sensor;
- the system can handle more users at the same time;
- no calibration nor training phases are required;
- privacy principles in the field of ubiquitous computing are followed, based on [6,19];
- a contribution to a computer vision problem (i.e. free gaze estimation) is given by our technique.

Preliminary experiments were conducted in our lab in order to quantitative validate the accuracy of the gaze estimation on different benchmarks of persons. Then the qualitative evaluation of the effectiveness of the proposed architecture was conducted in a real shop window demonstrating the ability to deal with real challenging environmental conditions. The followings of the paper is organized as follows: Sect. 2 gives an overview of the related works about gaze estimation and it highlights the contributions of the proposed solution with respect the leading state of the art methods. Section 3 deeply details the proposed method whereas in Sects. 4 and 5 the experimental setup and results are reported and discussed. Finally conclusions are in Sect. 6.

## 2 Related Works on Gaze Estimation

A survey of existing works and a detailed classification of methods can be viewed in [14]. Most gaze tracking methods are based on Pupil Center Corneal Reflection (PCCR) technique, [13,23]. It obtains the pose of the eye using the center of pupil contour and corneal reflections (glint) on the corneal surface from point light sources, usually one or multiple infrared (IR) lights. This method is not quite appropriate for general interactive applications. Usually a high-resolution camera is needed, and extra IR lights and the camera need to be calibrated carefully. Here, we concentrate on eye-tracking solutions without the usage of the beforehand exposed technique. These solutions can be divided into feature-based and appearance-based. Feature-based gaze estimation methods use extracted local features like contours or eye corners, while appearance-based methods utilize the image contents as an input with the intention of mapping these directly to screen coordinates, without requirements for scene geometry nor a camera calibration, but they need a significative high number of calibration point and, in general, are not head pose invariant. The work of [12] proposes an appearance-based method to achieve gaze estimation from multimodal Kinect data that is invariant to head pose, but it needs a learned person-specific 3D mesh model. In [8], after a one-time personal calibration, facial features are tracked and then used to estimate the 3D visual axis, proposing a 3D geometrical model of the eye. The method needs to accurately detect eye corners in order to create a complete 3D eye model. In [18], gaze tracking is performed using a stereo approach to detect the position and the orientation of the pupil in 3D space. A low-cost system with low-resolution webcam images, allowing for cursor control systems, is presented in [15]. The work of [16] proposes a method to estimate gaze tracking using a single and low-quality webcam. It limits head movements, but assumes that if the head moves, a head pose is estimated by an external program. A calibration phase is required. Valenti et. al [30] combine head pose and eye location information to accurately estimate gaze track. Eyes are located using isophote properties to obtain the center of semicircular patterns; head pose utilizes the cylindrical head model approach [34]. Their result are suitable for several applications, but they need a calibration phase and are tested only at a distance of 750 mm.

Most of state of the art work operates in constrained condition and needs a learning phase, using manually labeled data to train various type of classifiers. Furthermore, often a calibration phase occurs. Even methods that are considered unconstrained can work only in a very short range of head pose variations, and often the allowed translation is of less than 10 cm. Moreover, most of methods that produce the eye-gaze track are evaluated at a distance of 50–75 cm, when reported. Finally, often head-mounted devices are used.

Generally, head-pose is considered as a coarse gaze estimation technique. Authors of [28] assert that the head pose contributes to about 70 % of the visual gaze and focus of attention estimation based on head orientation alone can get an average accuracy of 88.7 % in a meeting application scenario. The perception of eye-gaze direction is also influenced by parameters like head contour and

nose angle [20]. Despite that, the idea of achieving gaze tracking using head pose information only is not new. A work that utilizes the same approach is in [7]. Here, using only the head pose information and given a model of the environment, the system is automatically able to give the estimation of the viewed object. In [27] a method for estimating where a person is looking in images where the head of a person is about 20 pixel high is presented. Here, eye information is not available, and estimation is made over head pose and the general body direction, combining direction and head pose using Bayes' rule to obtain the joint distribution over head pose and direction. The method proposed in [4] introduces the use of a classifier without any hand labelled data but based only on the output from an automatic tracking system in surveillance scenarios. In [11], a method that estimates the gaze direction accurately using information on both head and body pose directions and without using eye information is analyzed. Even in [3] the visual focus of attention is recognized by evaluating head pose information, getting anyway encouraging results. The work of [24] use head pose information to control a mouse, but investigating only the 2D information coming from a consumer camera. Finally, in [35], gaze direction is estimated by considering head posture information and using information that comes from the pupil, but it has been tested at a distance of 40 cm only.

Most of these methods pay attention only to the rough area of interest of the person, and are not seen as a possible technique to obtain the control of a device. Furthermore, no study is performed on the feasibility of an accurate gaze estimator that considers the more precise information that can be achieved using a device like a depth sensor. In the proposed work, a Microsoft Kinect is used to investigate same aspects, not only detecting the area of attention but trying also to achieve the exact position of gaze tracking ray.

In summary, our proposed work differs from the state of the art in the following aspects. First of all, we try to achieve the gaze estimation ray without using information different from head pose. Secondly, our method doesn't need both a training phase and a calibration phase. Thirdly, our work is tested with several different distance ranges going from 70 to 250 cm. Finally, in order to answer to our question, a full experimental setup was created and tested with different people, and all examinations and results are illustrated.

## 3 Proposed Method

### 3.1 Head Pose Estimation

The head pose estimation module takes care of supplying the information about rotation angle from the frontal pose, in terms of yaw, pitch and roll and translations, in meters, from the sensor's position.

Head pose estimation is a problem with 6-DOF, and can be represented with the parameter vector $\boldsymbol{p} = [\omega_x, \omega_y, \omega_z, t_x, t_y, t_z]$, where $\omega_x, \omega_y, \omega_z$ are the rotation parameters and $t_x, t_y, t_z$ are the translation parameters. They define the 3-DOF rotation matrixes $R_{3\times3}$ as:

$$R = \begin{bmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{bmatrix} \tag{1}$$

and the 3-DOF translation vector $T_{3\times1}$ as:

$$T = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{2}$$

The rigid motion of a head point $\boldsymbol{X} = [x, y, z, 1]^{\mathrm{T}}$ between time $t$ and time $t+1$ is:

$$\boldsymbol{X}(t+1) = M \times \boldsymbol{X}(t), \tag{3}$$

where M is defined as [21]:

$$M = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \tag{4}$$

Let point $\boldsymbol{X}(t)$ be projected on the image plane in $\boldsymbol{u} = \begin{bmatrix} u_x & u_y \end{bmatrix}^{\mathrm{T}}$. The explicit representation of the perspective projection function in terms of the rigid motion vector parameters and the coordinates of the point at $t+1$ is:

$$\boldsymbol{u}(t+1) = \begin{bmatrix} x - y\omega_z + z\omega_y + t_x \\ x\omega_z + y - z\omega_x + t_y \end{bmatrix} \cdot \frac{f_L}{-x\omega_y + y\omega_x + z + t_z}(t) \tag{5}$$

where $f_L$ is the focal length.

The system has been realized using the Kinect for Windows SDK [1]. The used model is the Candide-3 [2], and 2D coordinates of the key points on the aligned face in video frame coordinates are available. 121 2D feature points are tracked. By the Iterative Closest Point (ICP) [5] technique, by which a 3D point cloud model is iteratively aligned with the available 2D facial features (target), a 3D model on a detected face is built. The algorithm revises the transformation, i.e., combination of translation and rotation, needed to minimize the distance between the model and the target. Yaw, pitch and roll angles are extracted basing on the estimated rotation of the overlapped mask with respect to the Candide-3 model frontal pose. The X, Y, and Z position of the user's head are reported based on a right-handed coordinate system (with the origin at the sensor, Z pointed towards the user and Y pointed up).

For our purpose, both RGB and depth images are at a resolution of $640 \times 480$. Figure 1 shows the 3D mask overlapped to the 2D facial image in three different frames. From the figure it is possible to observe that the face tracker works also in presence of non frontal views. Multiple persons in the scene at the same time are also managed by the system.

### 3.2 Gaze Estimation

Our gaze estimation method works as follows. First of all, the 2D position of the detected eye center points are taken from the face mask. Note that small

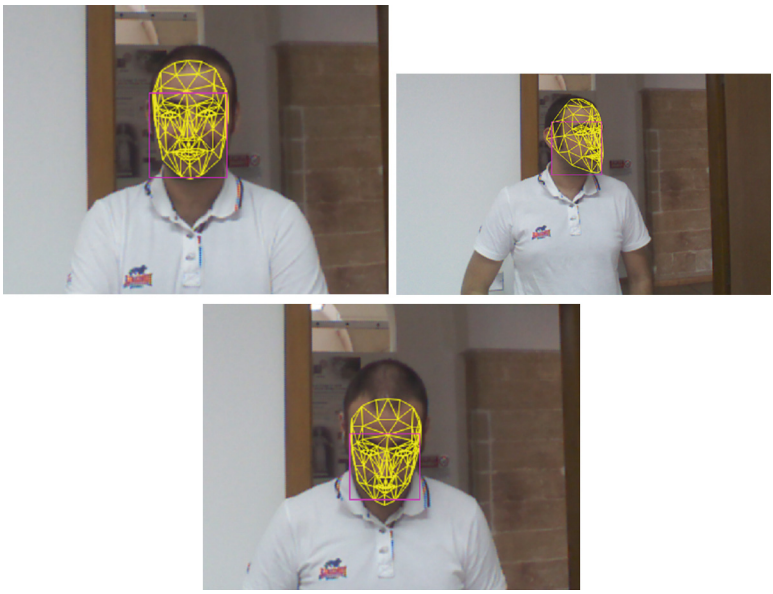


**Fig. 1.** Three different snapshots of the face tracking module.

occlusions are handled, and the eye center point is always estimated in the used model when the overlapping with the face successes. After that, the average value is taken, in order to take a point corresponding more or less with the nose septum and to use it as the origin of the gaze track. This value is converted into 3D coordinates with regard to a cartesian coordinate system centered inside the sensor. Starting from the head pose information, two angles are taken, i.e. $\omega_x$ and $\omega_y$, corresponding respectively to pitch and yaw. Then, the gaze track is computed and its intersection with a plane vertical with regard to the ground and passing from the center of the sensor is calculated. Note that with this method it is possible to achieve also the intersection point with every plane parallel to the considered one, just adding a translation parameter $k$ that will be algebraically added to the depth information, and then using the exposed procedure.

The intersection point is computed separately for each angle and using the same method, and the euclidean distance from the sensor can hereafter be computed. The procedure is showed for one angle in Fig. 2 and described in the followings. The Kinect sensor can give the length of the segment $\overline{AC}$ as the component $t_z$ of the translation vector $T$. It follows that, knowing a side and an angle, we can completely solve the right-angled triangle $\widehat{ABC}$. In particular, $\overline{AB} = \frac{\overline{AC}}{\cos \omega_y}$ and $\overline{BC} = \sqrt{\overline{AB}^2 - \overline{AC}^2}$. Using the same coordinate system, it is possible to compute also the cartesian equation of the gaze ray as the straight line passing for points $A = (x_A, y_A, z_A)$ and $B = (x_B, y_B, z_B)$ expressed as:

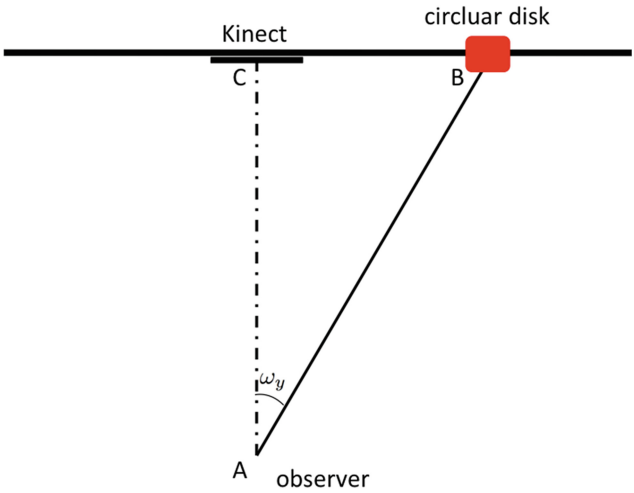



**Fig. 2.** A scheme of the gaze estimation solution.

$$r : \begin{cases} \frac{x-x_A}{x_B-x_A} = \frac{y-y_A}{y_B-y_A} \\ \frac{y-y_A}{y_B-y_A} = \frac{z-z_A}{z_B-z_A} \end{cases} \tag{6}$$

with $z_A = 0$ for the particular plane under consideration.

In case of translations on the $x$ and $y$ axes, the vector can be algebraical summed up with the computed value, in order to translate the gaze vector to the right position. Finally, in order to represent the real intersection point with the environment and to realize experimental tests, coordinates are normalized to image plane coordinates with the generic formula, valid for both coordinates $x$ and $y$ of the image plane:

$$c_{norm} = c - \frac{\text{bound}_{\text{low}}}{\text{bound}_{\text{upp}} - \text{bound}_{\text{low}}} \cdot \text{size}(I) \tag{7}$$

where $\text{bound}_{\text{upp}}$ and $\text{bound}_{\text{low}}$ are the two bounds, in meters, of the space, and $\text{size}(I)$ is the width (or height, depending on the coordinate in exam), expressed in pixels.

## 4 Experimental Setup

The environment for validation was defined as follows: a Microsoft Kinect device was positioned in front of the person, at a distance of 150 cm from the ground. Just behind the sensor, a panel with a set of 14 circular markers was positioned on its surface. Disks were distributed at a distance of 50 cm among $x$,$y$ or both axes. Figure 3 shows one quarter of the panel, exactly the upper-leftmost. All their distances from the sensor are known, and they are used as ground truth


**Fig. 3.** A portion of the used panel for testing.

information. Finally, the normalization Eq. 7 is used to bring back a distance range of $[-2\,\text{m}, 2\,\text{m}]$ into a square of extension of 1000 pixels. In Fig. 4, the reproduced window is showed. Here, the gaze point is automatically drawn as a red circle. Even the markers have been reported into the same image plane and are represented with a green circle; this way, the window realizes a possible cursor device. Even the small displacement between the sensor and the plane with ground truth data is managed by our system, using a $k$ value of 4 cm so that the total length of the side $\overline{AC}$ of the right-angled triangle will be $t_z + k$.

In both experimental setups, persons can be in the view angle of the sensor, i.e. 43 ° vertical by 57 ° horizontal field of view, and at a distance from the sensor in the range [40 cm, 300 cm] since we are operating in near mode (see Fig. 5 for more details). Head rotations are allowed in the three axes (also in the z-axes, because we are not using the Viola-Jones face detector), as also head translations.

All these working conditions are very suitable for a fully unconstrained system. Furthermore, using only the 3D information coming from the sensor to transform image coordinates into 3D camera coordinates and solving the gaze estimation as a three dimensional geometric problem, the system does not need any calibration phase. Finally, our algorithm has been tested with a frame rate of 30fps even on a common PC, i.e. an Ultrabook Intel i3 CPU @ 1.8 GHz with 4 GB of RAM, and was easily able to work in real-time. The usage of our technique on a common Ultrabook was made in order to facilitate for raw installations like in shelves or in wilder and nontraditional environments.

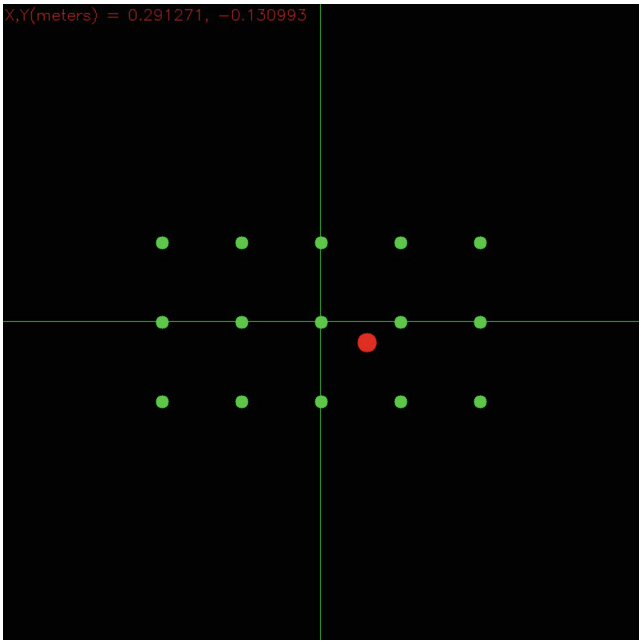



**Fig. 4.** The reproduced panel that realizes a cursor device.

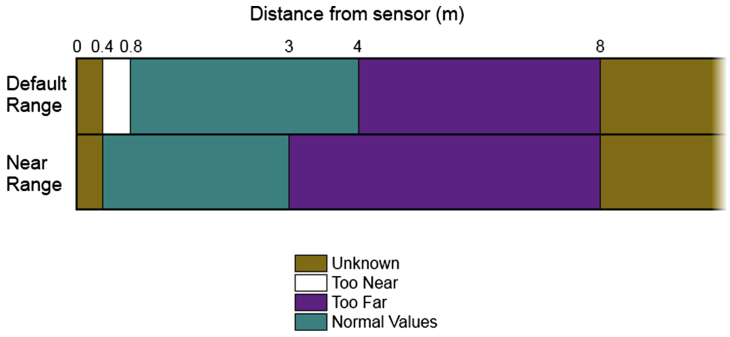



**Fig. 5.** The depth sensor has two depth ranges: the default range and the near range. The image shows the sensor depth ranges in meters.

## 5 Experimental Results

The proposed method has been tested with 9 different people. In order to get a comprehensive study, people are divided into three groups, three persons for each group. First group is composed by experienced people, i.e. people that know how the method works and that already tried the system before the test session. Second group is composed by people that try the system for the first time but that are well-informed about how the system works. Therefore, if they want to

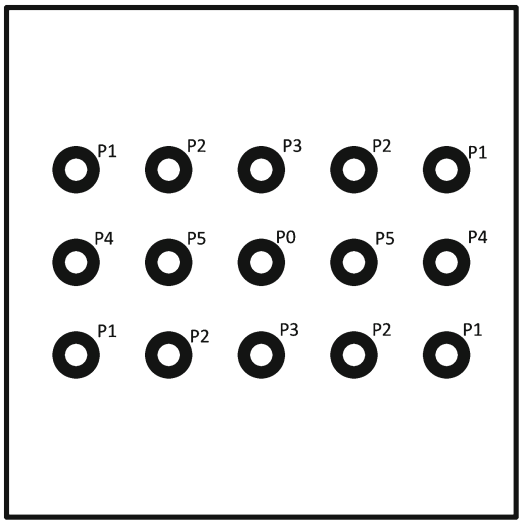



**Fig. 6.** The used grouping scheme for target points during tests.

point the attention on a given target, they will move the head in that direction without, for example, falling into the temptation to move only the eyes if the new point is very close to the previous one. Finally, in the last group there are unaware people that are just placed in front of the camera and are asked to point the markers. No constraints are given to the participants in terms of eyeglasses, beard or hairstyle and, in order to allow for wild settings, no panel or uniform background color has been put behind the participants.

The experiment is made as follows. People are asked to look at each of the placed markers, in a fixed order, using our real environments instead of a screen. Errors are measured as the angle between the ground truth gaze and the estimated gaze. Ground truth gaze is given by an oral feedback from the person, that stops moving when it is focused on a marker. For a given angle, the difference between the estimated gaze and the ground truth information increases if the distance grows. Differently from most of state of the art methods, in real environments the angle of error is considered also as a function that depends on the distance between the user and the sensor, because the error of the upstream method increases, unless the method is based on tools like head mounted devices. Even our head pose estimation tends to be inaccurate at the growing of the distance.

Results are showed in Tables 1, 2 and 3. Markers are divided into subset like in Fig. 6 in order to group together points that present the same distance from the sensor in terms of x, y or both axes, from P1 to P5, while P0 corresponds to the depth sensor position. For example, P4 are the points with a distance of 1 m from the sensor along the x axes and aligned along y axes, and so on. The second column shows the tested distance, i.e. 70, 150 and 250 cm. Errors are computed

**Table 1.** Experiments with the first group.

| | | Errors | | | |
|---|---|---|---|---|---|
| | | x (cm) | x (deg) | y (cm) | y (deg) |
| P0 | 70 cm | 1.50 | 1.22 | 2.66 | 2.18 |
| | 150 cm | 3.50 | 1.33 | 4.83 | 1.84 |
| | 250 cm | 6.00 | 1.37 | 8.50 | 1.94 |
| P1 | 70 cm | n.a. | n.a. | n.a. | n.a. |
| | 150 cm | 6.00 | 1.61 | 8.00 | 2.79 |
| | 250 cm | 2.60 | 0.52 | 4.00 | 0.88 |
| P2 | 70 cm | 8.77 | 5.03 | 7.66 | 4.37 |
| | 150 cm | 0.16 | 0.05 | 11.83 | 4.15 |
| | 250 cm | 4.33 | 0.95 | 5.83 | 1.29 |
| P3 | 70 cm | 5.61 | 4.58 | 3.50 | 1.94 |
| | 150 cm | 6.83 | 2.60 | 8.83 | 3.08 |
| | 250 cm | 4.66 | 1.06 | 1.83 | 0.40 |
| P4 | 70 cm | n.a. | n.a. | n.a. | n.a. |
| | 150 cm | 0.50 | 0.13 | 9.33 | 3.56 |
| | 250 cm | 0.66 | 0.13 | 15.66 | 3.47 |
| P5 | 70 cm | 3.66 | 2.03 | 3.83 | 3.13 |
| | 150 cm | 0.33 | 0.11 | 4.16 | 1.59 |
| | 250 cm | 4.33 | 0.95 | 8.16 | 1.87 |
| Total averages | 70 cm | 4.88 | 3.22 | 4.41 | 2.90 |
| | 150 cm | 2.88 | 0.97 | 7.83 | 2.83 |
| | 250 cm | 3.77 | 0.83 | 7.25 | 1.64 |

separately for each group and for yaw and pitch angle, in order to evidence where inaccuracies are located. In order to look at all the markers on the panel from the three positions, head pose of the users during the experiment with the three group can vary in the range $[-56.0\,°, +56.0\,°]$ in terms of yaw and in the range $[-35.5\,°, +35.5\,°]$ concerning pitch. Considering that a small angle from a wide distance corresponds to a bigger displacement, also errors in cm are reported. Note that "n.a." stands for a not available data, in our experiment exclusively due to the excessive rotation of the head to see objects at a far distance compared to the distance from the sensor, such that the Candide-3 model was not able to be overlapped on the face image from the system.

As can be observed, results for the first group are very accurate and, considering that all state of the art constraints are being removed, comparable to the state of the art methods in gaze estimation. Subsequently, the first group is perfectly able to control our device as it was a classic cursor device. The second group shows the same results, with some short outliers depending on the speed of the people to become familiar with the system. Anyway, they perform almost

**Table 2.** Experiments with the second group.

| | | Errors | | | |
|---|---|---|---|---|---|
| | | x (cm) | x (deg) | y (cm) | y (deg) |
| P0 | 70 cm | 2.50 | 2.04 | 2.33 | 1.90 |
| | 150 cm | 6.24 | 2.38 | 7.41 | 2.83 |
| | 250 cm | 28.5 | 6.50 | 13 | 2.97 |
| P1 | 70 cm | n.a. | n.a. | n.a. | n.a. |
| | 150 cm | 2.00 | 0.53 | 21.58 | 7.70 |
| | 250 cm | 19.00 | 3.84 | 27.00 | 6.05 |
| P2 | 70 cm | 10.16 | 5.89 | 17.16 | 10.40 |
| | 150 cm | 5.83 | 1.89 | 19.08 | 6.78 |
| | 250 cm | 3.00 | 1.50 | 0.65 | 0.33 |
| P3 | 70 cm | 2.83 | 2.31 | 8.33 | 4.77 |
| | 150 cm | 15.83 | 6.02 | 13.33 | 4.69 |
| | 250 cm | 18.5 | 4.23 | 20.5 | 4.58 |
| P4 | 70 cm | n.a. | n.a. | n.a. | n.a. |
| | 150 cm | 4.75 | 1.27 | 15.08 | 5.74 |
| | 250 cm | 33.00 | 6.79 | 12.50 | 2.86 |
| P5 | 70 cm | 11.16 | 0.83 | 6.51 | 0.68 |
| | 150 cm | 7.50 | 2.53 | 3.00 | 1.14 |
| | 250 cm | 22.50 | 5.03 | 19.00 | 4.34 |
| Total averages | 70 cm | 6.66 | 4.19 | 7.16 | 4.44 |
| | 150 cm | 6.98 | 2.44 | 13.25 | 4.81 |
| | 250 cm | 20.75 | 4.51 | 15.58 | 3.52 |

the same result as the first group, because the given information was easy to be assimilated. Even this group is able to use the device. The third group shows that some error can occur, but results are encouraging for applications where the focus of attention is the preponderant measure to be estimated. Finally, results are satisfactory even at a distance of 150 cm.

For informed users, during the experiments also the ability to monitor a possible device is tested in a dual way: first of all, the set of gaze points was registered and evaluated. The usage of the head pose information and the tracking algorithm applied with the facial mask model, has not shown outliers nor flickering effects. Finally, after each experiment, the virtual panel has been shown to the participants, and it was asked them to try to touch with the "gaze cursor" all the drawn fixed points from a distance of about 80–90 cm and to give a feedback. All of the participants felt comfortable and able to use our control device.

After validation, an intelligent shop window has been realized, as can be observed in Fig. 7. It has size of 3.70 m width and 2.80 m height. Between the user and the sensor, there is the window glass. This does not degrade overall

**Table 3.** Experiments with the third group.

| | | Errors | | | |
|---|---|---|---|---|---|
| | | x (cm) | x (deg) | y (cm) | y (deg) |
| P0 | 70 cm | 4.00 | 3.27 | 0.00 | 0.00 |
| | 150 cm | 15.00 | 5.71 | 10.00 | 3.81 |
| | 250 cm | 71.00 | 12.00 | 15.85 | 2.74 |
| P1 | 70 cm | n.a. | n.a. | n.a. | n.a. |
| | 150 cm | 17.00 | 4.73 | 29.00 | 10.46 |
| | 250 cm | 62.00 | 13.15 | 34.00 | 7.64 |
| P2 | 70 cm | 24.99 | 23.13 | 34.61 | 23.13 |
| | 150 cm | 1.78 | 0.61 | 10.90 | 3.82 |
| | 250 cm | 27.00 | 6.05 | 17.00 | 3.79 |
| P3 | 70 cm | 24.61 | 19.37 | 10.10 | 5.11 |
| | 150 cm | 33.20 | 12.48 | 17.10 | 6.06 |
| | 250 cm | 90.23 | 19.84 | 12.6 | 2.80 |
| P4 | 70 cm | n.a. | n.a. | n.a. | n.a. |
| | 150 cm | 3.90 | 1.01 | 4.80 | 1.83 |
| | 250 cm | 12.54 | 2.52 | 24.67 | 5.63 |
| P5 | 70 cm | 19.80 | 4.10 | 12.20 | 3.35 |
| | 150 cm | 2.78 | 0.96 | 17.38 | 6.60 |
| | 250 cm | 4.44 | 0.98 | 24.24 | 5.53 |
| Total averages | 70 cm | 18.35 | 12.67 | 12.20 | 7.90 |
| | 150 cm | 12.27 | 4.25 | 14.86 | 5.43 |
| | 250 cm | 44.61 | 9.75 | 20.75 | 4.69 |


**Fig. 7.** The realized shop window.

**Fig. 8.** The 2D projection of item occupancy in order to reveal the observed items.

performances, since Kinect generates IR light to determine an object's depth (distance) from the sensor, and this stream can pass over pane. Even due to the hardware under investigation, this system cannot work in outdoor environments, since the sensor cannot directly point towards direct sunlight.

To implement this system, the shop window has been reproduced on the screen. In order to define when an item should be considered observed by a user, the following method is used: considering Fig. 3, the gaze ray intersection with the virtual panel (the red circle) still continue to move in both horizontal and vertical direction inside an area that represent the shop window, using again Eq. 7. To each item of interest inside the shop window, a square area onto the screen, that represent a 2D projection of its space occupancy, has been manually defined. An item is considered observed if at least one point of the gaze ray is lying inside its square. Figure 8 shows our modeling. Finally, all results about gaze ray and observed items are stored. This way, data can be further used to realize decision making support system or to simply get useful stats, for example to detect most/least observed objects.

## 6 Conclusions

With this work, pervasive retail architecture based on a free gaze estimation system that can be used, for example, in a shop window to detect which items are observed from people has been proposed. The system exploits a depth sensor in order to extract head pose information, from which a fast geometric technique then evaluates the focus of attention of the persons in the scene (even more persons at the same time). Preliminary experiments were conducted in our lab in order to quantitative validate the accuracy of the gaze estimation on different benchmarks of persons. Then the qualitative evaluation of the effectiveness of

the proposed architecture was conducted in a real shop window demonstrating the ability to deal with real challenging environmental conditions. Future works will deal with the massive tests of the system in real shopping centers and on the definition of digital signage metrics for the exploitation of the extracted information for the definition of decision making support systems oriented to the creation/modification of the retail strategies.

## References


1. Apr 2014. http://msdn.microsoft.com/en-us/library/hh855347.aspx
2. Ahlberg, J.: Candide-3-an updated parameterised face (2001)
3. Ba, S.O., Odobez, J.M.: Recognizing visual focus of attention from head pose in natural meetings. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **39**(1), 16–33 (2009)
4. Benfold, B., Reid, I.: Unsupervised learning of a scene-specific coarse gaze estimator. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2344–2351. IEEE (2011)
5. Besl, P., McKay, N.D.: A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2), 239–256 (1992)
6. Brey, P.: Freedom and privacy in ambient intelligence. Ethics Inf. Technol. **7**(3), 157–166 (2005)
7. Brolly, X.L., Stratelos, C., Mulligan, J.B.: Model-based head pose estimation for air-traffic controllers. In: Proceedings of the 2003 International Conference on Image Processing, ICIP 2003, vol. 2, pp. II-113. IEEE (2003)
8. Chen, J., Ji, Q.: 3d gaze estimation with a single camera without ir illumination. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4. IEEE (2008)
9. van Doorn, M., van Loenen, E., de Vries, A.P.: Deconstructing ambient intelligence into ambient narratives: the intelligent shop window. In: Proceedings of the 1st International Conference on Ambient Media and Systems, p. 8. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2008)
10. Doshi, A., Trivedi, M.M.: Attention estimation by simultaneous observation of viewer and view. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 21–27. IEEE (2010)
11. Funatsu, N., Takahashi, T., Deguchi, D., Ide, I., Murase, H.: A study on gaze estimation using head and body pose information. In: International Workshop on Advanced Image Technology (2013)
12. Funes Mora, K., Odobez, J.M.: Gaze estimation from multimodal kinect data. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 25–30. IEEE (2012)
13. Guestrin, E.D., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Eng. **53**(6), 1124–1133 (2006)
14. Hansen, D., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010)
15. Ince, I.F., Kim, J.W.: A 2d eye gaze estimation system with low-resolution webcam images. EURASIP J. Adv. Signal Process. **2011**(1), 1–11 (2011)

16. Janko, Z., Hajder, L.: Improving human-computer interaction by gaze tracking. In: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), pp. 155–160. IEEE (2012)
17. Jellinger, K.: Cognitive processes in eye guidance. Eur. J. Neurol. **13**(9), e9 (2006)
18. Kohlbecher, S., Bardinst, S., Bartl, K., Schneider, E., Poitschke, T., Ablassmeier, M.: Calibration-free eye tracking by reconstruction of the pupil ellipse in 3d space. In: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, pp. 135–138. ACM (2008)
19. Langheinrich, M.: Privacy by design – principles of privacy-aware ubiquitous systems. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) UbiComp 2001. LNCS, vol. 2201, pp. 273–291. Springer, Heidelberg (2001)
20. Langton, S.R., Honeyman, H., Tessler, E.: The influence of head contour and nose angle on the perception of eye-gaze direction. Percept. Psychophysics **66**(5), 752–771 (2004)
21. Li, Z., Sastry, S.S., Murray, R.: A mathematical introduction to robotic manipulation (1994)
22. Mateo, J.C., San Agustin, J., Hansen, J.P.: Gaze beats mouse: hands-free selection by combining gaze and emg. In: CHI'08 Extended Abstracts on Human Factors in Computing Systems, pp. 3039–3044. ACM (2008)
23. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**(1), 4–24 (2005)
24. Nabati, M., Behrad, A.: Robust facial 2d motion model estimation for 3d head pose extraction and automatic camera mouse implementation. In: 2010 5th International Symposium on Telecommunications (IST), pp. 817–824. IEEE (2010)
25. Pieters, R.: A review of eye-tracking research in marketing. Rev. Mark. Res. **4**, 123–147 (2008)
26. Ravnik, R., Solina, F.: Audience measurement of digital signage: Quantitative study in real-world environment using computer vision. Interact. Comput. **25**(3), 218–228 (2013)
27. Robertson, N., Reid, I., Brady, J.: What are you looking at? gaze estimation in medium-scale images. In: Proceedings of the HAREM Workshop (in assoc. with BMVC) (2005)
28. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems, pp. 858–859. ACM (2002)
29. Tonkin, C., Ouzts, A.D., Duchowski, A.T.: Eye tracking within the packaging design workflow: interaction with physical and virtual shelves. In: Proceedings of the 1st Conference on Novel Gaze-Controlled Applications, p. 3. ACM (2011)
30. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. **21**(2), 802–815 (2012)
31. Voit, M., Stiefelhagen, R.: 3d user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, p. 51. ACM (2010)
32. Waizenegger, W., Atzpadin, N., Schreer, O., Feldmann, I., Eisert, P.: Model based 3d gaze estimation for provision of virtual eye contact. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 1973–1976. IEEE (2012)
33. Wetzel, P.A., Krueger-Anderson, G., Poprik, C., Bascom, P.: An eye tracking system for analysis of pilots' scan paths. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 1996. NTSA (1996)

34. Xiao, J., Moriyama, T., Kanade, T., Cohn, J.F.: Robust full-motion recovery of head by dynamic templates and re-registration techniques. Int. J. Imaging Syst. Technol. **13**(1), 85–94 (2003)
35. Zhang, W.Z., Wang, Z.C., Xu, J.K., Cong, X.Y.: A method of gaze direction estimation considering head posture. Int. J. Signal Process. Image Process. Pattern Recogn. **6**(2), 103–111 (2013)

# Face Re-Identification for Digital Signage Applications

Giovanni Maria Farinella[1(✉)], Giuseppe Farioli[1], Sebastiano Battiato[1], Salvo Leonardi[2], and Giovanni Gallo[1]

[1] Image Processing Laboratory, Department of Mathematics and Computer Science, University of Catania, Catania, Italy
{gfarinella,battiato,gallo}@dmi.unict.it, gfarioli88@gmail.com
http://iplab.dmi.unict.it
[2] Advice S.C., Catania, Italy
salvo.leonardi@adviceweb.it

**Abstract.** The estimation of soft biometric features related to a person standing in front an advertising screen plays a key role in digital signage applications. Information such as gender, age, and emotions of the user can help to trigger dedicated advertising campaigns to the target user as well as it can be useful to measure the type of audience attending a store. Among the technologies useful to monitor the customers in this context, there are the ones that aim to answer the following question: is a specific subject back to the advertising screen within a time slot? This information can have an high impact on the automatic selection of the advertising campaigns to be shown when a new user or a re-identified one appears in front the smart screen. This paper points out, through a set of experiments, that the re-identification of users appearing in front a screen is possible with a good accuracy. Specifically, we describe a framework employing frontal face detection technology and re-identification mechanism, based on similarity between sets of faces learned within a time slot (i.e., the models to be re-identified) and the set of face patches collected when a user appears in front a screen. Faces are pre-processed to remove geometric and photometric variability and are represented as spatial histograms of Locally Ternary Pattern for re-identification purpose. A dataset composed by different presentation sessions of customers to the proposed system has been acquired for testing purpose. Data have been collected to guarantee realistic variabilities. The experiments have been conducted with a leave-one-out validation method to estimate the performances of the system in three different working scenarios: one sample per presentation session for both testing and training (one-to-one), one sample per presentation session for testing and many for training (one-to-many), as well as considering many samples per presentation sessions for both testing and training (many-to-many). Experimental results on the considered dataset show that an accuracy of 88.73 % with 5 % of false positive can be achieved by using a many-to-many re-identification approach which considers few faces samples in both training and testing.


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 40–52, 2014.
DOI: 10.1007/978-3-319-12811-5_3

## 1 Introduction and Motivation

Digital signage is considered a revolutionary research area which aims to build advanced technologies for the out-of-home advertising. More specifically, with the term "digital signage" are referred the smart screens employed to show advertising content to a broad audience in a public/private area (e.g., store, airport, info office, taxi, etc.). The advertising screens are usually connected to the internet and are able to perform a series of "measurements" on the audience in front of the screen which are then exploited for marketing purposes (e.g., the screen reacts differently depending on the measurements). Taking into account the survey of the Aberdeen Research [1], the global market around the digital signage will expand from $1.3 billion in 2010 to almost $4.5 billion in 2016. The global market includes the displays, media players, as well as advanced software technologies for audience measurements. The only market of signage displays in 2014 is projected to reach more than 20 million units, with a growing in the years to come, and a total shipments hitting 25.8 million units by 2016 [2]. This rapid growth is due to the fact that digital signage gives to the organizations the possibility to advertise targeted and personalised messages to the audience in front of a smart screen. Thousands of organizations (e.g., retailers, government institutions, etc.) have already realized the benefits of the digital signage increasing the revenue related to their products or offering a better service in terms of given information to the audience.

In the context of digital signage, soft biometrics data inferred from the face of the user in front to an advertising screen [3] (such as gender identification and age estimation) are used to collect information to be exploited for users profiling. Ad-hoc advertising campaigns are then showed, taking into account of the collected information. Recent works demonstrate that computer vision techniques for face detection, age and gender recognition, classification and recognition of people's behavior can provide objective measurements (e.g., time of attention) about the people in front of a smart display [4–7]. Systems able to learn audience preferences for certain content can be exploited to compute the expected view time for a user, in order to organize a better schedule of the advertising content to be shown [5]. The audience emotional reaction can be also captured and analysed to automatically understand the feeling of the people to a campaign (e.g., to understand the attractiveness of a campaign with respect to another) [6]. Recent studies demonstrate that through computer vision methods it is possible quantify the percentage of the people who looked-at the display, the average attention time (differentiating by gender), the age groups who are more most responsive to the dynamic or static content [7].

Although the explosion of the field, in both academia and industry, it seems that measurements about the re-identification of a person in front a smart screen has been not taken into account in the context of digital signage. The information collected with a re-identification engine could be useful to answer the following question: is a specific person back to the advertising screen within a time slot? An automatic answer to this question obtained with a computer vision algorithm for the re-identification of person can be extremely useful to drive the

behaviour of the smart advertising screen which can automatically understand if the person is a new user or it has been seen already within a time slot. Looking at the humans' behaviour, the re-identification of a person is one of the most important feature used by the owner of stores to modify their behaviour in the presentation of the products or to propose special and personalised discounts. A huge number of digital signage scenarios can exploit the information about the re-identification. Among the others, the re-identification can be useful in giving personalised information at an ATM Bank and contextually can be exploited to improve the security during the withdrawal at the ATM (e.g., in a ATM session, the person who is taking the money should be the same of the re-identified one who has inserting the pin number).

In computer vision literature different methods for person re-identification have been proposed [9]. However, differently of the application contexts belonging to digital signage where in most of the cases only the person face is acquired to measure the information, the classic re-identification (e.g., in surveillance) is based on the exploitation of features extracted considering global appearance of an individual (e.g., clothing). Few works consider the re-identification based only on the person face. In [8] a re-identification method which include information of the face is proposed. The authors exploit face features jointly with other information (like hair, skin and clothes color) to re-identify a person.

In this paper, building on face recognition technologies, we propose and evaluate a re-identification system which works by considering only the face of the user in front an advertising screen. To this purpose we consider a dataset composed by $s = 100$ different presentation sessions (10 different customers per 10 different presentation to the system). Each session has been coupled with the remaining ones to produce a final set composed by $100 \times 99 = 9900$ different session pairs (training-session, testing-session) to be used for testing purposes. Data have been collected to guarantee variability during acquisition time (different acquisition period, geometric and photometric variability, faces appearance variability). Experimental results show that an accuracy of 88.73 % with 5 % of false positive can be achieved by using a many-to-many re-identification approach which considers few faces samples in both training and testing.

The reminder of this paper is organized as following. In Sect. 2 the proposed digital signage scenario is defined. In Sect. 3 the proposed approach is described, whereas experimental settings and results are reported in Sect. 4. Finally, Sect. 5 concludes the paper with hints for future works.

## 2 Proposed Digital Signage Scenarios

One of the key features to be a successful salesperson or a good front desk person in a info-point is the ability of identifying customers. In particular, it is well known that when a customer is re-identified it is more simple to offer the most appropriate products or information to the customer. A possible re-identification scenario is the one in which a person has asked some information about a cultural heritage place in a info point and comes back after few minutes

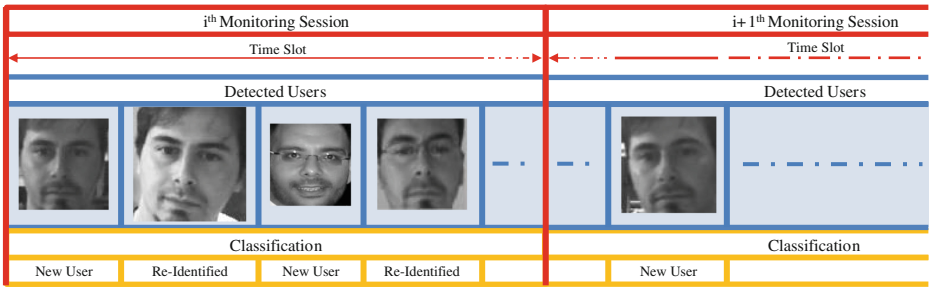



**Fig. 1.** A possible digital signage working scenario of a face re-identification engine. The system has to be able in re-identifying a customer within a monitoring session defined by a time slot.

to the desk. Depending on the information offered in the first discussion, the person at the front office can predict that some extra information are needed by the customer (e.g., the opening time, ticket price, etc.). This prior can help the front desk person being reactive to reply on further answers as well as in providing extra information to offer a better service. Similarly, if a customer has asked for a particular product (e.g., a perfume of a particular brand) and then comes back within a short time slot, there could be probability that he wishes to compare prices with respect to other products of the same type (e.g., in case he/she has not yet brought the previous seen products) or need to buy (or ask information about) other related products. A lot of other similar scenarios can be imagined. All of them, share the same customer's behaviour: he/she is back to the salesperson or to the front desk person within a time slot. In these cases the re-identification of the customer helps to be more effective in the service. Hence, independently from other collected information (name, age, gender, etc.) the re-identification information is useful.

The above scenarios can be straightforward translated in the context of Digital Signage: the smart screen has to be able to re-identify the customer in front to the display who has been seen within a time slot to trigger the right advertising campaign or to offer more information with respect to the ones searched previously by customer. Differently from the aforementioned scenarios in stores, in the case of Digital Signage the only information useful to perform the re-identification is usually the face of the customer which is visible by the camera of the smart screen.

In the considered context the Digital Signage system has to be designed to work as summarized in Fig. 1. During a monitoring session defined by a specific time slot, the system detects and classifies customers appearing in front of the screen as new or re-identified users. The re-identification engine should be robust to deal with both geometric (e.g., scale, orientation, point of view) and photometric (e.g., luminance) variabilities.

In the following sections we will detail all the key "ingredients" useful to build a Face Re-Identification engine. We will also report the experiments done

to assess the performances of the proposed system. We have tested the case of classifying the user as new or re-identified by setting a dynamic time slot equal to the time occurred from the previously seen customer. So, when a customer appears at the digital signage screen, it should be recognised if he is the same of the previous one or not. The results show that a re-identification accuracy of 88.73 % with 5 % of false positive can be obtained exploiting the current state of the art computer vision technologies.

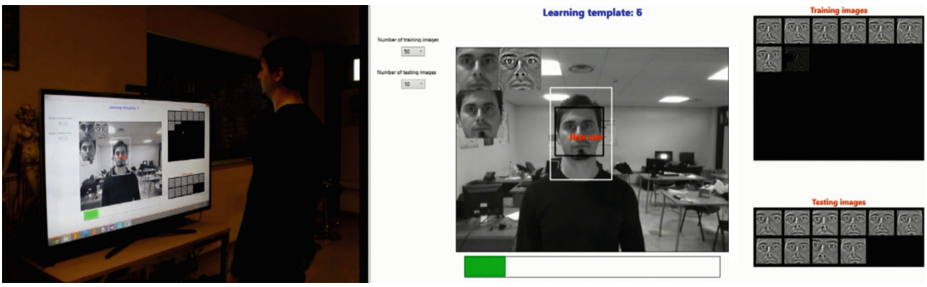



**Fig. 2.** The developed face re-identification engine in action.

## 3 Proposed Framework

In this section we detail all the component involved into the pipeline proposed for the face re-identification framework. As first, the face of the customer is detected (Fig. 3). Then the subimage containing the face is pre-processed to remove variabilities due to scale, rotation and lights condition changes (Figs. 4 and 5). From the obtained image the Local Ternary Patterns (LTP) [10] are extracted and the spatial-based distributions of LTP are considered as final representation of the detected face. For the re-identification purpose (Fig. 6), we employ the $\chi^2$ distance between a set of $N$ representations obtained considering $N$ frames in which the face of the customer is detected, and a set of $M$ representations related the face of the previously seen customer. We use the Kinect [11] as acquisition system by exploiting the skeleton tracking library to detect and track the customer. The aforementioned face re-identification pipeline is performed exploiting the Kinect's RGB channel in a region surrounding the skeleton's head position.

### 3.1 Face Detection

As first stage of the Face Re-Identification pipeline the face of the user has to be detected. To this purpose we exploit the well-known Viola and Jones object detection framework [12]. It is able of processing images in real time achieving high face detection rates. The solution exploits the "Integral Image" representation so that Haar-like features can be computed at any scale or location in constant time. A learning algorithm based on AdaBoost [13] is used to select

the most discriminative features bases for classification purposes. Combining different classifiers in a "cascade" the background regions of the image are discarded while faces are detected. In our experiments we constrain the Face Re-Identification to frontal-faces so that both eyes are visible into the detected face. This is useful to have references points (i.e., the eyes) to align the detected faces hence removing variabilities due to scale and rotation. As for the face detection, the eyes are localized through the Viola and Jones framework (Fig. 3).

### 3.2 Face Pre-processing

When a frontal face containing both eyes is detected, a pre-processing step is performed to prepare image containing the face for the feature extraction. As first, geometric variability due rotation and scale changes are removed by mean of the parameters obtained trough an affine transformation of the detected eyes coordinates with respect to fixed eyes positions (Fig. 4). Then to counter the photometric variabilities of illumination, local shadowing and highlights, the normalization pipeline suggested in [10] is employed (Fig. 5). The pipeline is composed by the following three ordered steps: gamma correction, difference of gaussian (DoG) filtering and contrast equalization. All the details can be found in [10].

### 3.3 Feature Extraction and Face Representation

At this stage the faces have been detected and aligned. As in the problem of Face Recognition [14], for the Face Re-Identification the faces have to be represented with discriminative descriptors to deal with facial expression, partial occlusions and other changes. Moreover, the signature of a face should be computed in real-time. Different papers addressing the problems related to the measurement of soft biometrics (e.g., gender, ages, etc.) for digital signal applications exploit features which are variants of the so called Local Binary Patterns (LBP) [15–17]. LBP are robust to lighting effects because they are invariant to monotonic gray-level transformations, and are powerful in discriminating textures. The LBP is an operator useful to summarise local gray-level structures. It takes a local neighborhood around a pixel $p$, performs a thresholding of the pixels of the neighborhood by considering the value of the pixel $p$ and uses the resulting binary-valued patch as a local image descriptor. By considering a neighborhoods of $3 \times 3$ around $p$, LBP gives a 8 bit code (i.e., only 256 possible codes). The codes extracted for each pixel of the patch containing the face can be then summarized in a normalised histogram and used as final representation for soft biometric measurements. As a drawback LBP tends to be sensitive to noise, especially in near-uniform image (as in the case of facial regions). For this reason in our system we employ a generalization of the LBP, the so called Local Ternary Patterns (LTP) [10]. Differently than LBP where the central pixel $p$ is considered as threshold to obtain a binary code, the LTP operator gives three possible values as output for each neighbor of $p$. Let $p'$ a neighbor of $p$ and $t$ a user-specified threshold[1]. The LTP operator is

[1] In our experiments the LTP threshold $t$ has been fixed as suggested in [10].

**Fig. 3.** Customer's face detected by the face re-identification framework. The face is considered for further processing only if both eyes are detected. Eyes position are used to remove scale and rotation variabilities before feature extraction.

**Fig. 4.** Top row: Customer's face detected in different frames of a session by the Face Re-Identification framework. Bottom Row: Corresponding faces after rotation and scale alignment based on eyes positions.

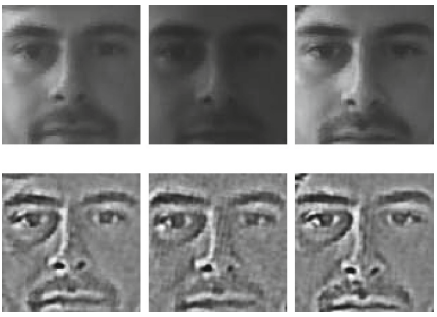


**Fig. 5.** Top row: Customer's face detected in different frames of a session and aligned by the Face Re-Identification framework. Bottom Row: Corresponding faces after photometric normalization [10].

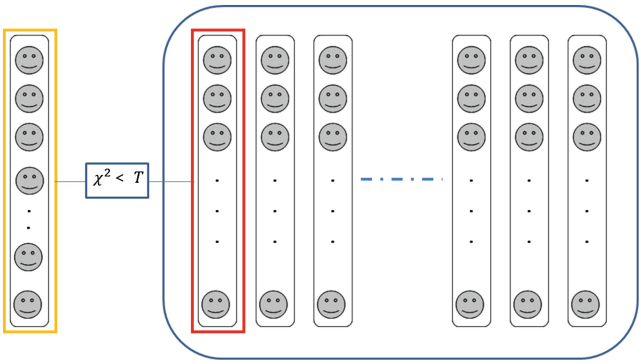



**Fig. 6.** For the re-identification purpose we compute the $\chi^2$ distance between a set of $N$ representations obtained considering $N$ frames in which the face of the customer is detected (in orange), and a set of $M$ representations related the face of the last seen customer (in red) (Color figure online).

defined as follows:

$$LTP(p, p', t) = \begin{cases} 1 & \text{if } p' \geq p + t \\ 0 & \text{if } |p' - p| < t \\ -1 & \text{if } p' \leq p - t \end{cases} \quad (1)$$

For the encoding procedure the LTP code is splitted into positive and negative patterns which are considered unsigned to compute two different binary codes (i.e., Upper Pattern and Lower Pattern). For representation purposes, the two distributions of Upper and Lower Patterns are computed and concatenated.

As suggested in other works related the field of soft biometrics estimation [18], we divide the input image with a regular grid ($7 \times 7$ in our experiments). The LTP distributions are computed on each cell of the grid to encode spatial information and hence improve the face re-identification results.

### 3.4 Face Re-Identification

As last step the framework re-identifies a customer within a monitoring session defined by a given time slot (Fig. 1). We consider the case in which the time slot is equal to the amount of time in which the last customer has been detected by the digital signage system. This means that the re-identification engine has to be able to answer the following question: is the current customer and the previous seen customer the same person? The proposed face re-identification mechanism works as described in the following. During a customer session (i.e., the customer is interacting with the smart display) the face re-identification engine collects $M$ faces templates as described in Sects. 3.1 and 3.2, and represents them as described in the previous section. This set of templates is considered the training dataset for the re-identification of the next customer. When the next customer

shows up in front the smart monitor, the proposed system analyses its face, collects and represents $N$ faces templates (testing images) and finally compares them with the $M$ training images learned during the previous monitoring session (Fig. 6). To this purpose the $\chi^2$ distance is employed. For every represented template of the current customer, the smallest $\chi^2$ distance to any represented template of the training is computed. Then the smallest distance is considered to decide if the current customer and the previous one are the same person (re-identification) or if there is a new user. The decision threshold is fixed to have a low number of false positives (i.e., low number of cases in which a customer $B$ is wrongly re-identified as previous seen customer $A$ rather than as new user). The number of training templates could be greater than the number of testing template (i.e., $M > N$). Indeed, although the collection of training images of a new user does not have impact to the system behaviour and it is transparent for the customer in front the digital screen and can have duration equal to the time spent by the customer in front the screen, during testing the system have to react in real time (or at least as soon as he get the $N$ faces templates) so that personalised information can be offered to the final user. Our experiments demonstrates that few templates can be used for re-identification purpose in both training ($M = 50$) and testing ($N = 6$).

### 3.5 Customer Tracking

In real scenarios of digital signage it is reasonable to imagine that the customer is surrounded by other people (which look or not at the smart screen). To properly track the customer we use the Kinect and its standard skeleton tracking. This allows to recognize the closest person to the screen which we assume to be the customer interacting with the advertising screen. All the processing needed for the face re-identification (see previous section) is performed exploiting the Kinect's RGB channel. Since we track the skeleton, the customer's head position is known and the face re-identification pipeline can be performed only in the region surrounding the head of the customer (i.e., white bounding box in Fig. 2).

## 4 Experimental Settings and Results

To asses the performances of the proposed framework we have collected a dataset composed by 5000 faces. Specifically, we have acquired 100 different presentation sessions (each composed by 50 face patches) related to 5 male and 5 female different users who have been presented at the system 10 different times. The data have been acquired at different time (in a range of a couple of months), in different environments to guarantee variability in the subjects' appearance (beard, makeup, tan), as well as by considering different photometric and geometric conditions (lights, point of view, clutter background). Each presentation session has been coupled with the remaining ones to produce a final set of $100 \times 99 = 9900$ users' pairs (training-session, testing-session) to be used in the experiments. Each pair of users has been labeled as a session containing the same person or not.

In each session, a customer showed up in front the smart screen and a RGB-D video of a couple of minutes has been acquired with a Kinect device [11]. From each session the first $M = 50$ face templates have been detected, processed and represented as described in previous section. For testing purpose we have used a leave one out approach [19]. At each run, the templates related to one customer presentation session are considered as training model, whereas the remaining sessions templates are considered as testing set. Different tests have been done to assess the influence of the size $M$ of the set of faces template to be used as training in a session, and the size $N$ of the set of faces to be used as testing in a session. The final results are obtained by averaging over all runs. Since the face re-identification results depends from the threshold used to make the re-identification decision (see Sect. 3.4) we evaluate the proposed method making use of ROC curves [19].

In Fig. 7 are reported the ROC curves results obtained at varying of the parameter $M$ related to the faces templates to be considered as model in the training set. In this test the number of faces template to be used for testing is fixed to $N = 1$. In Fig. 8 are reported the ROC curves results obtained at varying of the parameter $N$ related to the number of faces template to be considered as testing set. The number of templates to be used for training is instead fixed to $M = 50$. The related results are compared in Fig. 9. The tests showed that by increasing the number of patterns to be considered for both training and test sets the face re-identification performances improve. It should be noted that for the

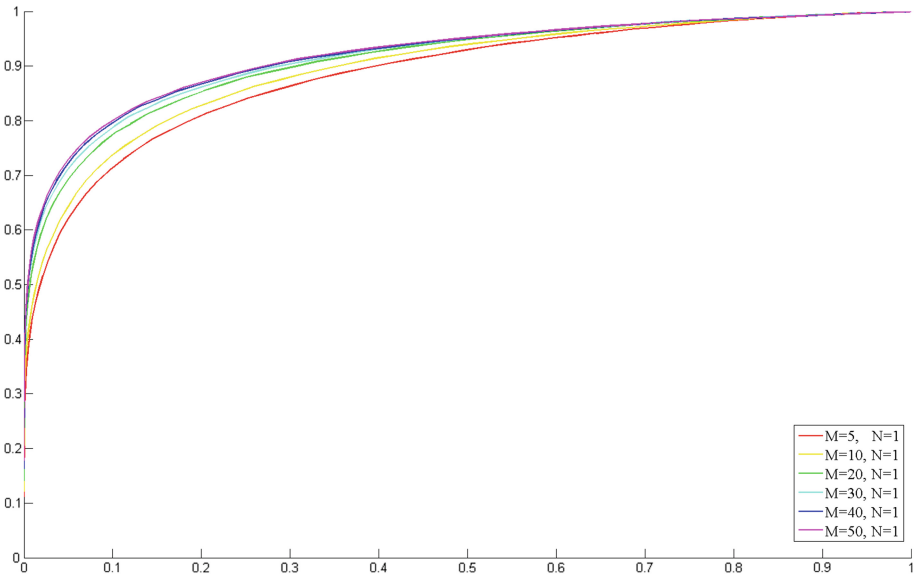



**Fig. 7.** Face re-identification results at varying of the number of training templates. The vertical axis corresponds to the True Positive rate, whereas the horizontal axis is related to the False Positive rate.

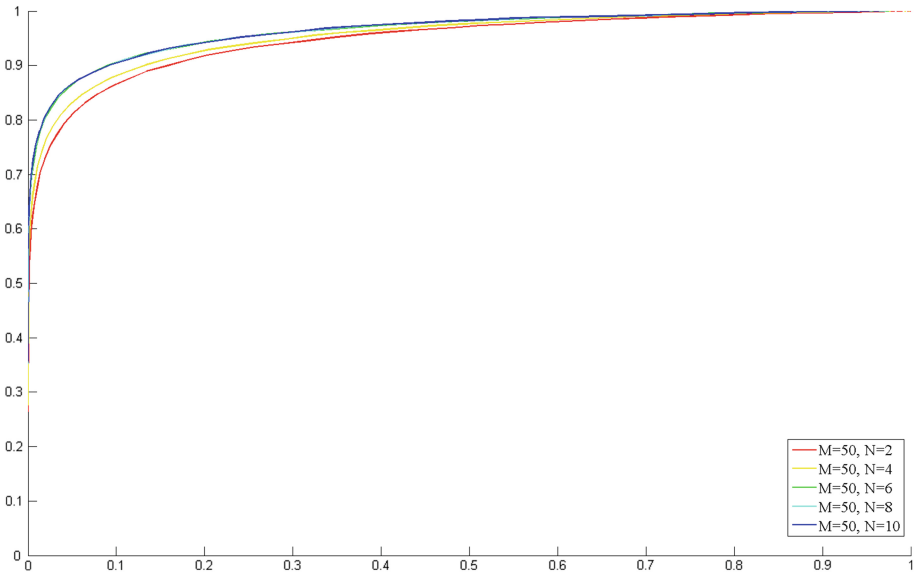



**Fig. 8.** Face re-identification results at varying of the number of testing templates. The vertical axis corresponds to the True Positive rate, whereas the horizontal axis is related to the False Positive rate.

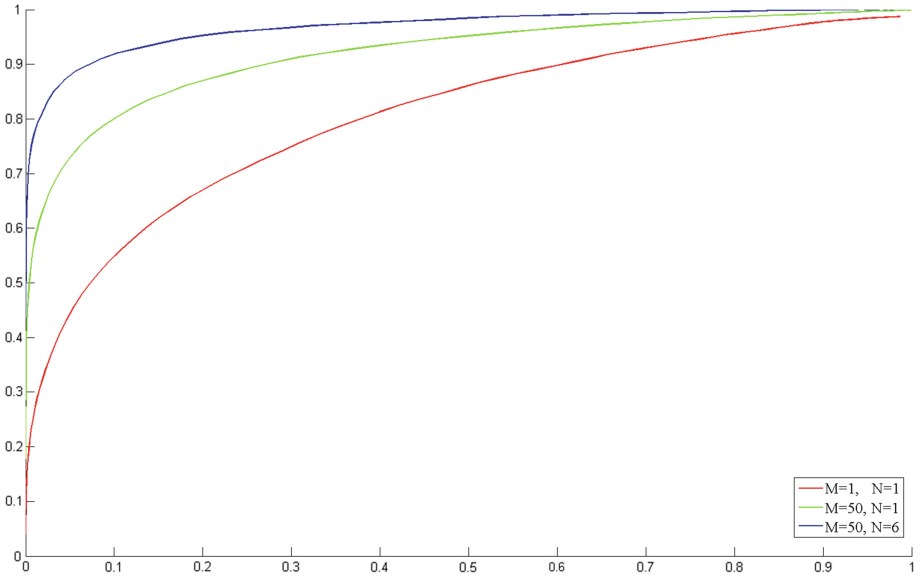



**Fig. 9.** Comparison of face re-identification results at varying of the number of training and testing templates. The vertical axis corresponds to the True Positive rate, whereas the horizontal axis is related to the False Positive rate.

testing case (i.e., when a user shows up to the screen and should be recognized as new user of same person of the previous customer) by increasing the number of patterns more than 6 per session does not gives much improvements in the results (Fig. 8). Considering few face templates in both training ($M = 50$) and testing ($N = 6$) the re-identification accuracy stated at 88.73 % with 5 % of false positive. The proposed face re-identification framework (Fig. 2) works in real time as demonstrated by the video at the following link: http://iplab.dmi.unict.it/download/VAAM2014.

## 5 Conclusions and Future Works

This paper has addressed the problem of face re-identification in the context of digital signage applications. Motivations, scenarios and the main ingredients to build a face re-identification framework are described. Quantitative evaluation is reported to assess the proposed framework. Further works will be devoted to improve the performances of the proposed baseline face re-identification method by reviewing and refining all the steps involved into the face re-identification pipeline and augmenting the face representation to consider depth information (e.g., for faces alignment purpose). Moreover, an in-depth study by considering the re-identification of the last $K > 1$ customers on a larger dataset collected in real scenarios (e.g. stores, info points) should be done.

## References

1. Aberdeen Research: Digital Signage: A Path to Customer Loyalty, Brand Awareness and Marketing Performance (2010)
2. Khatri, S.: Digital signage and professional display market set for solid growth in 2012. Signage & Professional Displays Market Tracker Report, April 2012
3. Ricanek, K., Barbour, B.: What are soft biometrics and how can they be used? Computer **44**(9), 106–108 (2011)
4. Batagelj, B., Ravnik, R., Solina, F.: Computer vision and digital signage. In: Tenth International Conference on Multimodal Interfaces (2008)
5. Müller, J., Exeler, J., Buzeck, M., Krüger, A.: ReflectiveSigns: digital signs that adapt to audience attention. In: Tokuda, H., Beigl, M., Friday, A., Brush, A.J.B., Tobe, Y. (eds.) Pervasive 2009. LNCS, vol. 5538, pp. 17–24. Springer, Heidelberg (2009)
6. Exeler, J., Buzeck, M., Müller, J.: eMir: digital signs that react to audience emotion. In: 2nd Workshop on Pervasive Advertising, pp. 38–44 (2009)
7. Ravnik, R., Solina, F.: Audience measurement of digital signage: quantitative study in real-world environment using computer vision. Interact. Comput. **25**(3), 218–228 (2013)
8. Dantcheva, A., Dugelay, J.-L.: Frontal-to-side face re-identification based on hair, skin and clothes patches. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance, pp. 309–313 (2011)
9. Vezzani, R., Baltieri, D., Cucchiara, R.: People re-identification in surveillance and forensics: a survey. ACM Comput. Surv. **46**(2), 29:1–29:3 (2013)

10. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process. **19**(6), 1635–1650 (2010)
11. Microsoft Kinect. http://www.microsoft.com/en-us/kinectforwindows/
12. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
13. Schapire, R.E.: A brief introduction to boosting. In: International Joint Conference on Artificial intelligence (1999)
14. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. **34**(4), 399–485 (2003)
15. Ahonen, T., Hadid, A., Pietiknen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
16. Wang, J.-G., Yau, W.-Y., Wang, H. L.: Age categorization via ECOC with fused gabor and LBP features. In: Workshop on Applications of Computer Vision (2009)
17. Hadid, A., Pietikinen, M.: Combining appearance and motion for face and gender recognition from videos. Pattern Recognit. **42**(11), 2818–2827 (2009)
18. Ylioinas, J., Hadid, A., Pietikainen, M.: Age classification in unconstrained conditions using LBP variants. In: International Conference on Pattern Recognition (2012)
19. Webb, A.R.: Statistical Pattern Recognition, 2nd edn. Wiley, New York (2002)

# Evaluation of LBP and HOG Descriptors for Clothing Attribute Description

Javier Lorenzo-Navarro(✉), Modesto Castrillón, Enrique Ramón, and David Freire

Instituto Universitario SIANI, Campus Universitario de Tafira, Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas, Spain
javier.lorenzo@ulpgc.es

**Abstract.** In this work an experimental study about the capability of the LBP, HOG descriptors and color for clothing attribute classification is presented. Two different variants of the LBP descriptor are considered, the original LBP and the uniform LBP. Two classifiers, Linear SVM and Random Forest, have been included in the comparison because they have been frequently used in clothing attributes classification. The experiments are carried out with a public available dataset, the *clothing attribute dataset*, that has 26 attributes in total. The obtained accuracies are over 75 % in most cases, reaching 80 % for the necktie or sleeve length attributes.

## 1 Introduction

In most cultures, clothing is related to gender, age or social status. In video analytics demographics is a key element, so the identification of clothes can be used as an additional cue providing information about the consumer style or interest. The clothing identification problem can be divided into three categories. On the one hand, there exists the problem of clothes segmentation whose aim is to segment the image into regions that corresponds to the same garment. Another problem is the clothing attribute classification where the garment is described by its attributes as color, pattern, neck type, sleeve and others. The last problem is the clothing recognition where the objective is to detect garment categories in the images as t-shirt, dress, trousers and so. This paper focuses on the problem of clothing attribute classification.

The use of clothing to improve other identification task is not new. Satta et al. [12] propose a variation of their Multiple Component Matching framework named Multiple Component Dissimilarity (MCD) in other to compute the dissimilarity between a prototype person built using textual description of clothing attributes and those that made up the gallery. Gallagher and Chen [6] compute color and texture features for each superpixel obtained from a previous

J. Lorenzo: Work partially funded by the Institute SIANI and the Departamento de Informática y Sistemas at ULPGC.


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 53–65, 2014.
DOI: 10.1007/978-3-319-12811-5_4

segmentation process and finally the clothing mask is calculated. In an application of image retrieval, Borras et al. [1] propose a method for describing upper body clothing making use of texture and color features. The different parts of the clothes are obtained with a split-and-merge approach using homogeneity measures as criteria. Also in content-based image retrieval, Weber et al. [14] introduce a novel approach by getting the mask of the clothing starting from a set of trained pose detectors in order to deal with occlusions and different poses inherent to humans. Manfredi et al. [8] present an approach for segmenting garments in fashion stores databases. As features, color and HOG are used and combined with a Gaussian Mixture model. On the other hand, a method based in computing the color and texture of the clothing regions obtained from a background subtraction phase is described by Yang and Yu [16]. Clothing has even used in estimating the human occupation in images [13].

Recently some works have been proposed to describe clothing based on attributes. In this regard, Yamaguchi et al. [15] describe clothes based on a method which labels superpixels obtained from a previous segmentation process making use of a Conditional Random Field model. The method starts from a pose estimation that is re-evaluated using the clothing predictions. In a recommendation scenario, Kalantidis et al. [7] describe an approach for segmenting and recognizing clothing. They start from a segmentation of the person and then each segment is classified by computing the LSH index. A similar approach is introduced by Chen et al. [4] but the user assists in the segmentation process. Liu et al. [11] have a proposal for describing clothes which is based on pose estimation as [15] and using as features color, SIFT and HOG are able to classify clothes into 23 categories. A method for classifying upper body clothing using Random Forests where decisions in each node include a SVM to estimate the density around each decision boundary is described in [2]. Chen et al. [3] present a method for describing clothing by attributes. They define 11 attributes which are classified using 40 features obtained from the arms and torso. For each attribute a SVM is learned and the CRF is used in order to take into account the relation that exists among some attributes as collar and necktie.

The aim of this paper is to study the ability of two well known descriptors along with color to classify clothing attributes as sleeve length, the existence of collar, fabric pattern and color. In this work we only focus on upper body clothing. The paper is organized as follows. In Sect. 2, a brief description of the LBP and HOG descriptor will be given. Section 3 describes the attributes that are considered in this work and the features used for classifying each one. In Sect. 4, the results obtained for the different attributes under study are shown and compared in two different experimental setups. Finally in Sect. 5, the conclusions are presented.

## 2 LBP and HOG Descriptors

In this section a brief description of the Local Binary Patterns (LBP) and the Histogram of Oriented Gradients (HOG) is given. LBPs have been used successfully in different Computer Vision problems since its original application to

texture classification [10]. In their definition, each pixel is encoded taking into consideration its neighborhood by means of a threshold. LBPs are therefore easily computed, and have shown their capacity of discrimination in different real world problems, while exhibiting a notorious robustness to monotonic gray-scale changes. In the original definition by Ojala et al. [9], the operator labels each image pixel comparing it with its $3 \times 3$ neighborhood, and encodes the pixel as a binary number. The resulting codes are used as a histogram to represent the texture. This definition has recently been extended to a set of arbitrary circular neighborhoods. The expression to compute the generalized LBP is the following:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{1}$$

where, $g_c$ is the gray level of $P$ neighbors, $g_p(p = 0, 1, \cdots, P-1)$. The function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{2}$$

An extension of the LBP is the uniform LBP, $LBP^{u2}$. A LBP code is considered uniform if the binary pattern contains at most two bitwise transitions, from 0 to 1 or vice versa when the bit pattern is considered circular. For example, patterns of 00000000 (0 transitions), number 00110000 to (2 transitions) and 11100111 (2 transitions) are uniform, while the patterns of 11001001 (4 transitions) and 01001010 (6 transitions) are not uniform. Uniform patterns indicate that some patterns are more frequent than others. Among the 256 original LBPs using the 8 neighbors with radius 1, just 58 of them are uniform, and the rest are non uniform, there are a total number of 59 different labels considering uniform LBPs.

HOG encloses a histogram in its definition [5], as the gradient orientations of a regular area are represented by means of a histogram. A whole image is represented by the concatenation of the histograms of the collection of regular areas, i.e. cells. The illumination influence is reduced normalizing each cell considering its neighborhood, called block. In their work, Dalal et al. made use of a cell size of $8 \times 8$ pixels, and the block is $2 \times 2$ cells. We have assumed that implementation, restricting the number of bins to 9. For the normalization stage we have used L1.

## 3 Attribute Clothing Classification

Pose estimation is an important element in clothing attribute segmentation and classification. Some authors have used the state of the art pose estimator [17] but in several situations the correct pose is not retrieved. So in other to avoid the influence of a bad pose estimation, in this work we have annotated manually the dataset. As we focus on upper body clothing, 16 annotation points are defined for each person: 2 for the face, 4 for each arm and 6 for the torso (Fig. 1).

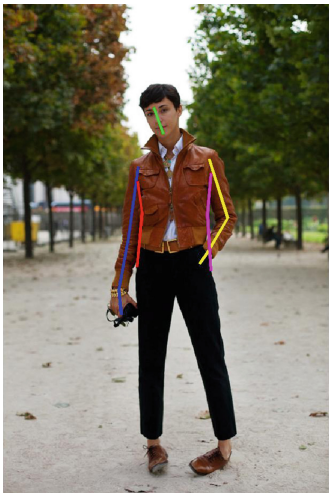


**Fig. 1.** Manual annotation of the upper body: face (green), right arm (blue), left arm (yellow), right torso (red) and left torso (magenta) (Color figure online).

### 3.1 Sleeve Length

The hypothesis in this work is that computing the descriptors in those segments can give more information for the attribute in study. So, once the different parts of the body are obtained the sleeve length attribute is related to the portion of arm that is covered by fabric. In order to detect the amount of exposed skin in the arms there is the possibility of training a skin detector but it has to face up with different illumination and color skin. Here, a different approach is proposed. Normally sleeves are made of the same fabric than the rest of the upper body clothing, so we compare the color histogram of each arm part (hand, arm and forearm) with the color histogram of the torso. According to this, a six element feature vector is defined as

$$x_{arm\_segment} = diff(color\,histogram_{torso}, color\,histogram_{arm\_segment}) \quad (3)$$

where $arm\_segment \in \{right\,hand, right\,arm, right\,forearm, left\,hand, left\,arm, left\,forearm\}$ and $diff$ is the Bhattacharyya distance between the two histograms.

### 3.2 Collar and Necktie

In order to classify the existence of collar and necktie the region of interest to consider is the neck and the upper chest. Considering the annotations of the images, this region can be obtained as the rectangle with base located in the armpits and height in the chin (Fig. 3). To compute the HOG descriptor in this region, it is previously normalized to $32 \times 24$ pixels and then HOG is computed

using a cell size of $8\times8$ pixels, block size of $16\times16$ pixels, 9 bins for orientation and $8\times8$ pixels overlapping between blocks. The two variants of the LBP descriptor, original and uniform, are computed on the original rectangle obtaining 256 and 59 bin histograms respectively. The descriptors are obtained from the histograms after a $\ell_1$ normalization stage.

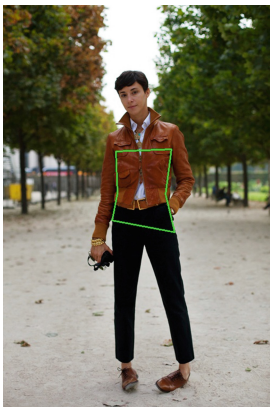


**Fig. 2.** Torso region obtained from the annotations.

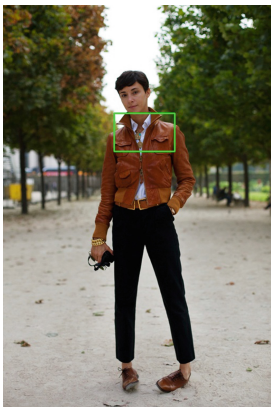


**Fig. 3.** Neck region obtained from the annotations.

### 3.3 Fabric Pattern

The pattern of the fabric is mainly defined by the torso region so this region will be used to compute the features used to classify the different patterns. The torso region is defined by six points: 2 for armpits, 2 for waist and 2 for hips (Fig. 2). The histogram of the LBP in this region is computed and it is used as feature vector for pattern classification.

### 3.4 Color

The last attribute we have considered in this work is the color of the clothes. The histogram of the CIE L*a*b* is computed for the torso (Fig. 2) and it is used as feature for color classification.

## 4 Experiments

The experiments has been carried out using the *clothing attribute dataset* [3] which has 1856 images. It has 26 attributes in total, including 23 binary-class attributes: 6 for pattern, 11 for color and 6 miscellaneous; and 3 multi-class attribute: sleeve length, neckline shape and clothing category.

Some attributes exhibit a very unbalanced class distribution where a majority classifier can yield a high accuracy. For example the class distribution for

**Table 1.** Results for necktie, collar and sleeve length attributes with the balanced dataset with 10-CV.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Necktie (HOG) | **82.47 %** | 0.83 | 0.83 | 0.83 | 80.48 % | 0.79 | 0.85 | 0.82 |
| Necktie (LBP) | 72.60 % | 0.74 | 0.74 | 0.74 | 65.41 % | 0.64 | 0.74 | 0.69 |
| Necktie ($LBP^{u2}$) | 67.12 % | 0.69 | 0.67 | 0.68 | 68.49 % | 0.68 | 0.75 | 0.71 |
| Collar (HOG) | **78.40 %** | 0.78 | 0.78 | 0.78 | 74.35 % | 0.72 | 0.79 | 0.75 |
| Collar (LBP) | 66.91 % | 0.68 | 0.64 | 0.66 | 60.04 % | 0.58 | 0.67 | 0.62 |
| Collar ($LBP^{u2}$) | 65.61 % | 0.69 | 0.55 | 0.61 | 59.67 % | 0.58 | 0.69 | 0.63 |
| Sleeve Length (RGB) | 70.28 % | 0.70 | 0.70 | 0.70 | 70.60 % | 0.69 | 0.74 | 0.71 |
| Sleeve Length (HSV) | **77.83 %** | 0.78 | 0.78 | 0.78 | 77.64 % | 0.79 | 0.79 | 0.79 |

**Table 2.** Results for necktie, collar and sleeve length attributes training with the balanced dataset and testing with the unbalanced dataset.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Necktie (HOG) | 84.02 % | 0.96 | 0.84 | 0.88 | **86.39 %** | 0.99 | 0.86 | 0.92 |
| Necktie (LBP) | 76.02 % | 0.99 | 0.76 | 0.86 | 73.32 % | 0.97 | 0.74 | 0.84 |
| Necktie ($LBP^{u2}$) | 71.60 % | 0.99 | 0.71 | 0.83 | 74.71 % | 0.98 | 0.75 | 0.85 |
| Collar (HOG) | **78.76 %** | 0.80 | 0.79 | 0.79 | 71.97 % | 0.55 | 0.80 | 0.65 |
| Collar (LBP) | 69.70 % | 0.53 | 0.66 | 0.59 | 63.85 % | 0.47 | 0.74 | 0.57 |
| Collar ($LBP^{u2}$) | 69.37 % | 0.53 | 0.56 | 0.54 | 60.50 % | 0.44 | 0.72 | 0.54 |
| Sleeve Length (RGB) | 75.62 % | 0.87 | 0.76 | 0.78 | 71.08 % | 0.17 | 0.71 | 0.27 |
| Sleeve Length (HSV) | **81.62 %** | 0.90 | 0.82 | 0.85 | 78.74 % | 0.23 | 0.79 | 0.35 |

yellow color is 67 positive samples and 1677 negative samples so a majority classifier achieves a 96 % of accuracy. To avoid this, for each attribute the samples were divided into two datasets. The balanced dataset where the classes has a balanced distribution, and the rest of the samples in another dataset, the unbalanced dataset. For example, for the yellow attribute the balanced dataset has 58 positive samples and 57 negative samples, and the unbalanced dataset has 19 positive samples and 1620 negative samples.

The results were obtained in two different settings for the training of the classifier. One experiment was to test the accuracy of the classifier with a 10-fold cross validation using the balanced dataset. The other experiment was to train the classifier with the balanced dataset and test it with the unbalanced dataset. For the experiments a linear SVM classifier with C = 5 and a Random Forest were used.

Tables 1 and 2 show the results for the necktie, collar and sleeve length for the two experimental settings: 10-fold CV with the balanced dataset and train with the balanced dataset and test with the unbalanced dataset. It can be observed that for the necktie attribute the highest accuracy, 82.47 %, is obtained with

**Table 3.** Accuracy for necktie and collar attributes combining HOG and LBPs descriptors.

| | 10-CV setup | | Train/Test setup | |
|---|---|---|---|---|
| | SVM Linear | Random Forest | SVM Linear | Random Forest |
| Necktie (HOG+LBP) | **82.88 %** | **82.88 %** | **85.21 %** | **85.21 %** |
| Necktie (HOG + $LBP^{u2}$) | 80.48 % | 79.11 % | 85.28 % | 84.80 % |
| Collar (HOG + LBP) | **78.07 %** | 74.53 % | 78.03 % | 74.46 % |
| Collar (HOG + $LBP^{u2}$) | 77.30 % | 73.99 % | **78.90 %** | 73.05 % |

**Table 4.** Results for pattern attribute using HOG with the balanced dataset with 10-CV.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Floral | **86.49 %** | 0.91 | 0.83 | 0.87 | 81.08 % | 0.77 | 0.92 | 0.84 |
| Graphics | **81.76 %** | 0.81 | 0.86 | 0.83 | 78.24 % | 0.76 | 0.86 | 0.81 |
| Plaid | **69.05 %** | 0.70 | 0.72 | 0.71 | 68.45 % | 0.68 | 0.76 | 0.72 |
| Solid | 70.75 % | 0.70 | 0.74 | 0.72 | **75.82 %** | 0.74 | 0.80 | 0.77 |
| Spot | 71.26 % | 0.75 | 0.69 | 0.72 | **74.85 %** | 0.73 | 0.84 | 0.78 |
| Stripe | 62.72 % | 0.64 | 0.62 | 0.63 | **68.86 %** | 0.65 | 0.86 | 0.74 |
| Average | 73.67 % | 0.75 | 0.74 | 0.75 | 74.55 % | 0.72 | 0.84 | 0.78 |

the SVM classifier with the balanced dataset and with the Random Forest, 86.39 %, in the second setting. Furthermore, the F-measure is above 0.80 for both classifiers. With both classifiers, HOG descriptors exhibits a better performance than LBP and $LBP^{u2}$ than can be explained for the typology of the necktie attribute that is defined by straight lines.

The results for the collar attribute show lower accuracy than for the necktie attribute with an accuracy of 78 % in both settings for the SVM classifier and in both cases the accuracy is higher than the one obtained with the Random Forest. Moreover, the presence of collar is evidenced with clear contours that are better captured by the HOG descriptors than by the LBP one due to their different nature.

In order to test if LBP and HOG descriptors encode complementary information, an experiment using as descriptor the combination of the LBP and HOG in only one descriptor is carried out. Table 3 shows the results obtained in this experiments and except in one case the accuracy is very similar or even lower than using only HOG as features.

The features for the sleeve length attribute were computed using two different color spaces in Eq. (3): RGB and HSV. In Tables 1 and 2 can be observed the improvement of the performance when the HSV color space is used to compute the difference between the arm segments and the torso region with both classifiers. On the other hand, for HSV, accuracies of 70.28 % and 81.62 % are

**Table 5.** Results for pattern attribute using HOG training with the balanced dataset and testing with the unbalanced dataset.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Floral | 76.29 % | 1.00 | 0.76 | 0.86 | **88.43 %** | 1.00 | 0.89 | 0.94 |
| Graphics | **85.76 %** | 0.99 | 0.86 | 0.92 | 85.20 % | 0.99 | 0.86 | 0.92 |
| Plaid | 67.43 % | 0.99 | 0.68 | 0.80 | **80.28 %** | 0.99 | 0.81 | 0.89 |
| Solid | **75.71 %** | 0.35 | 0.78 | 0.49 | 67.08 % | 0.29 | 0.83 | 0.43 |
| Spot | 68.19 % | 0.99 | 0.68 | 0.81 | **83.45 %** | 0.99 | 0.84 | 0.91 |
| Stripe | 60.24 % | 0.98 | 0.60 | 0.75 | **79.18 %** | 0.99 | 0.80 | 0.88 |
| Average | 72.27 % | 0.88 | 0.73 | 0.77 | 80.60 % | 0.88 | 0.84 | 0.83 |

**Table 6.** Results for pattern attribute using LBP with the balanced dataset with 10-CV.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Floral | 77.98 % | 0.79 | 0.78 | 0.77 | **78.38 %** | 0.78 | 0.83 | 0.81 |
| Graphics | 78.82 % | 0.79 | 0.79 | 0.79 | **84.71 %** | 0.83 | 0.89 | 0.86 |
| Plaid | **77.38 %** | 0.78 | 0.77 | 0.77 | 75.00 % | 0.74 | 0.82 | 0.78 |
| Solid | 73.53 % | 0.74 | 0.74 | 0.74 | **77.45 %** | 0.75 | 0.84 | 0.79 |
| Spot | **78.44 %** | 0.78 | 0.78 | 0.78 | 71.86 % | 0.71 | 0.80 | 0.75 |
| Stripe | **74.56 %** | 0.76 | 0.74 | 0.74 | 68.42 % | 0.66 | 0.79 | 0.72 |
| Average | 76.79 % | 0.77 | 0.77 | 0.77 | 75.97 % | 0.75 | 0.83 | 0.79 |

obtained for the two experimental settings with SVM classifier. Note that the performance of the classifier does not decrease in both experimental setups. For RGB the performance is lower but the behavior is similar to the HSV, increasing the accuracy in the second experimental setting.

Tables 4 and 5 present the results obtained in the pattern attribute classification using HOG descriptors. It can be observed that Random Forest performs better than SVM because in both experimental setups the average accuracy is higher. In the second experimental setup the difference in performance is more emphasized with 80.60 % for Random Forest versus 72.27 % for SVM.

The results for the different LBP descriptors that were considered in the classification of the fabric pattern attribute, original LBP and the $LBP^{u2}$, are shown in Tables 6, 7, 8 and 9. Tables 6 and 7 show the results obtained with the original LBP descriptor for each pattern. The accuracy using SVM classifier for all patterns is around the 77 % except for the *Solid* and *Stripe* patterns that drop to 73.53 % and 74.56 %. In the second experimental setup with the same classifier, the performance is worse except for the *Spot* and *Stripe* patterns

**Table 7.** Results for pattern attribute using LBP training with the balanced dataset and testing with the unbalanced dataset.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Floral | 71.56 % | 0.99 | 0.71 | 0.82 | **78.16 %** | 1.00 | 0.78 | 0.88 |
| Graphics | 73.45 % | 0.98 | 0.73 | 0.83 | **88.00 %** | 1.00 | 0.88 | 0.94 |
| Plaid | 76.65 % | 0.98 | 0.76 | 0.85 | **85.50 %** | 0.99 | 0.86 | 0.92 |
| Solid | **75.26 %** | 0.85 | 0.75 | 0.78 | 71.17 % | 0.32 | 0.82 | 0.45 |
| Spot | **86.73 %** | 0.98 | 0.87 | 0.92 | 83.26 % | 0.99 | 0.83 | 0.91 |
| Stripe | **87.00 %** | 0.97 | 0.87 | 0.91 | 84.44 % | 0.99 | 0.85 | 0.91 |
| Average | 78.44 % | 0.96 | 0.78 | 0.85 | 81.76 % | 0.88 | 0.84 | 0.84 |

**Table 8.** Results for pattern attribute using $LBP^{u2}$ with the balanced dataset with 10-CV.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Floral | 74.77 % | 0.77 | 0.75 | 0.75 | **77.48 %** | 0.81 | 0.77 | 0.79 |
| Graphics | 70.00 % | 0.70 | 0.70 | 0.70 | **77.06 %** | 0.76 | 0.82 | 0.79 |
| Plaid | **75.00 %** | 0.76 | 0.75 | 0.75 | 73.21 % | 0.73 | 0.78 | 0.75 |
| Solid | 73.37 % | 0.74 | 0.74 | 0.73 | **73.69 %** | 0.71 | 0.83 | 0.76 |
| Spot | **76.05 %** | 0.76 | 0.76 | 0.76 | 71.26 % | 0.71 | 0.79 | 0.74 |
| Stripe | **73.68 %** | 0.76 | 0.74 | 0.75 | 71.26 % | 0.71 | 0.79 | 0.74 |
| Average | 73.81 % | 0.75 | 0.74 | 0.74 | 73.99 % | 0.74 | 0.80 | 0.76 |

that yield higher accuracy. When Random Forest is used as classifier, results have shown a higher variance because there are better accuracies as the one for *Graphics* with 84.71 % but also there are worse accuracies like 68.42 % for *Stripe* and 71.86 % for *Spot*. The same behavior is observed when the Random Forest is trained with the balanced dataset and tested with the unbalanced one. On average the SVM classifier is better than the Random Forest classifier. However, in a multiple bi-class problem like this, a solution can be the combination of classifiers depending on the pattern, so if the best classifier for each pattern is considered we obtain on average an accuracy of 78.49 %.

The results for the pattern attribute using the $LBP^{u2}$ descriptor are shown in Tables 8 and 9 respectively. As can be seen, the performance for the SVM classifier is worse than with the LBP descriptor, being the accuracy lower for all patterns except for the *Graphics* pattern in the second experimental setup. It is shown that the use of the Random Forest classifier increases the performance in both experimental settings for most of the patterns. Considering both classifiers it can be observed than the Random Forest classifier performs better than the

**Table 9.** Results for pattern attribute using $LBP^{u2}$ training with the balanced dataset and testing with the unbalanced dataset.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Floral | 65.53 % | 0.98 | 0.65 | 0.78 | **82.45 %** | 1.00 | 0.82 | 0.90 |
| Graphics | 76.55 % | 0.98 | 0.77 | 0.85 | **82.96 %** | 0.99 | 0.83 | 0.91 |
| Plaid | 76.40 % | 0.98 | 0.76 | 0.85 | **79.52 %** | 0.99 | 0.80 | 0.88 |
| Solid | 69.24 % | 0.84 | 0.96 | 0.73 | **71.17 %** | 0.31 | 0.81 | 0.45 |
| Spot | 79.97 % | 0.98 | 0.80 | 0.87 | **82.94 %** | 1.00 | 0.83 | 0.91 |
| Stripe | **88.94 %** | 0.97 | 0.88 | 0.92 | 83.68 % | 0.99 | 0.84 | 0.91 |
| Average | 76.11 % | 0.96 | 0.80 | 0.83 | 80.45 % | 0.88 | 0.82 | 0.83 |

**Table 10.** Results for color attribute with the balanced dataset with 10-CV.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Brown | **80.00 %** | 0.80 | 0.80 | 0.80 | 76.89 % | 0.74 | 0.84 | 0.79 |
| Red | 75.18 % | 0.77 | 0.75 | 0.75 | **85.11 %** | 0.84 | 0.89 | 0.86 |
| Yellow | 81.90 % | 0.82 | 0.82 | 0.82 | **90.48 %** | 0.91 | 0.91 | 0.91 |
| Green | 76.03 % | 0.76 | 0.76 | 0.76 | **79.34 %** | 0.78 | 0.82 | 0.80 |
| Cyan | 77.46 % | 0.77 | 0.77 | 0.77 | **80.92 %** | 0.82 | 0.86 | 0.84 |
| Blue | **79.55 %** | 0.79 | 0.79 | 0.79 | **79.55 %** | 0.80 | 0.87 | 0.83 |
| Purple | **76.52 %** | 0.77 | 0.77 | 0.77 | 72.17 % | 0.71 | 0.80 | 0.75 |
| White | 67.43 % | 0.68 | 0.67 | 0.67 | **71.43 %** | 0.69 | 0.75 | 0.72 |
| Gray | **73.97 %** | 0.74 | 0.74 | 0.74 | 71.58 % | 0.69 | 0.77 | 0.73 |
| Black | 79.00 % | 0.79 | 0.79 | 0.79 | **79.34 %** | 0.78 | 0.82 | 0.80 |
| Average | 75.66 % | 0.76 | 0.76 | 0.76 | 75.83 % | 0.75 | 0.81 | 0.78 |

SVM one and if the best accuracy for each pattern is considered as in the previous case, the average accuracy increases to 84.44 %.

Tables 10 and 11 show the results of the accuracy obtained with the color attribute. As it was said in Sect. 3.4, the histograms of the CIE L*a*b* color space have been used as features. With both classifiers in the first experimental setup, the accuracy for this attribute does not have a similar behavior. On the one hand, there is a color, *Yellow*, with an accuracy 90.48 % for the Random Forest classifier. On the other hand, for the same setting we observe an accuracy of 67.43 % for the *White* color with the SVM classifier. In the first experimental setup both classifiers perform similar on average and if only the best classifier for each color is considered the accuracy reaches until 79.67 %. When the classifiers are trained with the balanced dataset and tested with the unbalanced one, the

**Table 11.** Results for color attribute training with the balanced dataset and testing with the unbalanced dataset.

| | Linear SVM | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F-measure | Acc. | Prec. | Rec. | F-measure |
| Brown | 76.77 % | 0.96 | 0.77 | 0.84 | **82.09 %** | 0.98 | 0.83 | 0.90 |
| Red | 80.91 % | 0.98 | 0.81 | 0.88 | **89.39 %** | 1.00 | 0.89 | 0.94 |
| Yellow | 79.44 % | 0.99 | 0.79 | 0.87 | **85.60 %** | 1.00 | 0.86 | 0.92 |
| Green | 83.55 % | 0.98 | 0.84 | 0.90 | **87.74 %** | 1.00 | 0.88 | 0.93 |
| Cyan | 76.38 % | 0.99 | 0.76 | 0.86 | **86.31 %** | 1.00 | 0.86 | 0.93 |
| Blue | **88.65 %** | 0.96 | 0.89 | 0.92 | 86.94 % | 0.99 | 0.88 | 0.93 |
| Purple | 70.23 % | 0.98 | 0.70 | 0.81 | **80.36 %** | 0.99 | 0.81 | 0.89 |
| White | 72.37 % | 0.84 | 0.72 | 0.77 | **78.04 %** | 0.95 | 0.80 | 0.86 |
| Gray | 73.50 % | 0.90 | 0.75 | 0.79 | **79.81 %** | 0.96 | 0.82 | 0.88 |
| Black | 79.15 % | 0.86 | 0.79 | 0.81 | **82.26 %** | 0.94 | 0.83 | 0.88 |
| Average | 76.71 % | 0.92 | 0.77 | 0.83 | 82.29 % | 0.97 | 0.83 | 0.90 |

Random Forest behaves better than SVM although the same variations in accuracy are observed in this experimental setting, ranging from 70.23 % for *Purple* with SVM to 89.93 % for *Red* with Random Forest. Considering again the use of a combination of classifiers with the best one for each color, the average accuracy increases 84.05 %.

## 5 Conclusions

In this work, LBP and HOG descriptors and color have been used for clothing attribute identification obtaining good results. The experiments were carried out with a public dataset, the *clothing attribute dataset*, to allow other researchers to compare their results with those published in this paper. Instead of using a pose estimator, the dataset images were manually annotated to avoid the influence of the bad pose estimations in the results that can hinder the actual discriminative power of the descriptors. For each individual, the head, torso and arms were annotated.

Some of the five attributes that have been considered in this work exhibits a very unbalanced class distribution in the dataset so we have defined two experimental setups. The first one with a dataset with equal proportion for each class and the second one with the rest of samples which results in an unbalanced distribution. Two of the most used classifier in clothing recognition field were compared, SVM and Random Forest. When an attribute, as pattern and color, has several classes neither of them performs better than the other for all the attributes so our conclusion is that the use of a combination of classifier will be the preferred option. With the combination of classifier higher accuracies are obtained, reaching 84.05 % and 84.44 % for color and pattern respectively.

For necktie, collar and sleeve length attributes the best accuracies were 86.39 %, 78.76 % and 81.62 % respectively.

## References


1. Borras, A., Tous, F., Llados, J., Vanrell, M.: High-level clothes description based on colour-texture and structural features. In: 1st Iberian Conference on Pattern Recognition and Image Analysis IbPRIA (2003)
2. Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Gool, L.V.: Apparel classification with style. In: 11th Asian Conference on Computer Vision. Daejeon, Korea, November 5–9 2012
3. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012)
4. Chen, J.L., Chen, W.Y., Chen, I.K., Chi, C.Y., Chen, J.L.: Interactive clothing retrieval system. In: IEEE International Conference on Consumer Electronics (ICCE) (2014)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) International Conference on Computer Vision and Pattern Recognition, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, vol. 2, pp. 886–893, June 2005
6. Gallagher, A., Chen, T.: Clothing cosegmentation for recognizing people. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
7. Kalantidis, Y., Kennedy, L., Li, L.J.: Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In: International Conference on Multimedia Retrieval (ICMR) (2013)
8. Manfredi, M., Grana, C., Calderara, S.: A complete system for garment segmentation and color classification. Mach. Vis. Appl. **25**(4), 955–969 (2014)
9. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**, 51–59 (1996)
10. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns. Springer, Heidelberg (2011)
11. Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., Yan, S.: Fashion parsing with weak color-category labels. IEEE Trans. Multimedia **16**, 253–265 (2014)
12. Satta, R., Pala, F., Fumera, G., Roli, F.: Person Re-Identification. In: Chapter People Search with Textual Queries About Clothing Appearance Attributes, pp. 371–390. Springer, Heidelberg (2014)
13. Song, Z., Wang, M., Hua, X.S., Yan, S.: Predicting occupation via human clothing and contexts. In: IEEE International Conference on Computer Vision (ICCV) (2011)
14. Weber, M., Buml, M., Stiefelhagen, R.: Part-based clothing segmentation for person retrieval. In: International Conference on Advanced Video and Signal-based Surveillance (AVSS) (2011)
15. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

16. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: 18th IEEE International Conference on Image Processing (2011)
17. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)

# Features Descriptors for Demographic Estimation: A Comparative Study

Pierluigi Carcagnì, Marco Del Coco(✉), Pier Luigi Mazzeo, Andrea Testa, and Cosimo Distante

CNR-INO, Arnesano, Italy
marco.delcoco@ino.it

**Abstract.** Estimation of demographic information from video sequence with people is a topic of growing interest in the last years. Indeed automatic estimation of audience statistics in digital signage as well as the human interaction in social robotic environment needs of increasingly robust algorithm for gender, race and age classification. In the present paper some of the state of the art features descriptors and sub space reduction approaches for gender, race and age group classification in video/image input are analyzed. Moreover a wide discussion about the influence of dataset distribution, balancing and cardinality is shown. The aim of our work is to investigate the best solution for each classification problem both in terms of estimation approach and dataset training. Additionally the computational problem it considered and discussed in order to contextualize the topic in a practical environment.

## 1 Introduction

Last years saw the increase of signage digital flat-panel displays in public places providing the distribution of commercial advertisement and multimedia content [13,21]. The agility of these systems to provide different and attractive advertising content is widely used by company to influence customer's behavior. Anyway the variety of people present in typical public places used for digital signage distribution proposes a more attractive challenge: the ability to interact, among time, with people present and specific place in a specific moment. A wide range of possible interaction has been actually analyzed including touch, gesture and speech [26,29]. Anyway this kind of approach requests an active interaction of people rarely reachable due to shyness or haste. This display blindness [28] effect can be avoided by means of a passive audience measurement. Audience measurement in digital signage is a field of growing interest in last years especially for marketing purposes. Indeed a statistical knowledge of the audience observing a panel is a key factor for company that could adapt marketing strategies ongoing without the risk approach to wait for sales results. Moreover advanced systems could also provide an adaptative content scheme, so that the systems can modify the advertise just looking to the typology of people interested in the panel in the meanwhile. A group of teenagers could cause the generation of a console game advertise. To the opposite a woman could be more interested


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 66–85, 2014.
DOI: 10.1007/978-3-319-12811-5_5

in cosmetics and health-care products or a Asian guy in typical Oriental food restaurants. A future application could be oriented to the smart-TV too.

Social Robots is another field strictly related with the automatic extraction of demographics information. Social robots are automatic or semi-automatic robots able to interact with human-being following basic social rules. The use of these kind of machines is of high interest for applications in interactions and study of children with autism [31] or in older adults assistance [35]. All of these contexts need the capacity of the robots to recognize who is the subject of the interaction (Male-Female, Young-Old) and adapt its behavior to the specific situation. Anyway this kind of robots does not provide high computational and memory resources so that the selection of robust as well light classification algorithms is a key goal.

Moreover these information are likewise interesting in forensic and video surveillance applications.

Computer vision techniques allow to estimate some demographics information as gender, race and age, from video sequences of a monocular camera. Anyway, extraction of these kind of information is not trivial due to ambiguity related to the single person anatomy and lifestyle. A similar discussion can be done for race recognition where the somatic aspect of some population could be not well defined and where some people member of a population could exhibit aspect of another one.

All these problems have been well investigated during last years in computer vision and machines learning fields. Well known Viola and Jones approach introduces a robust cascade detector (based on AdaBoost [14] and Haar features) for the face recognition in images [41] actually considered as a state-of-art approach.

Gender is probably the easiest issue. It is a two class classification problem. Mäkinen and Raisamo [27] and Sakarkaya et al. [33] present two wide interesting survey that exhaustively treat the topic. First study was done by Brunelli and Poggio [5] (1995) that investigate the use of geometrical features and Abdi et al. [15] (1995) that applied pixel based methods. Lyons et al. used Gabor wavelets with PCA and Linear discriminant analysis (LDA) [25]. In 2002 Sun et al. show the importance of features selection for generic algorithms [37] and successively test the efficiency of LBP for gender classification [36]. Saatci and Town applied Active Appearance Models (AAM) to this scope [32] with the support of an SVM classifier. Recently Ullah et al. show the performance of a spatial Weber Local Descriptor (SWLD) [40]. Other recent works as [7] use facial and non facial features (Uniform LBP and HOG) and, for classification, a stacking of classifiers, each one focused in a particular family of features. In [34] a procedure to learn discriminative LBP-Histogram (LBPH) bins, as compact facial representation for gender classification, is presented. Classification is performed by adopting SVM with the selected LBPH bins. In [38] the authors focus on fusion and features selection methods based on mutual information as a measure of relevance and redundancy among features. Another key point is the dependence of the prediction algorithm by uncontrollable factors as head pose and age, as presented in [2]. More specifically the authors show the

difference in performance for different age categories and highlight the fragility of face alignment. Finally, [11] discusses the importance of a cross validation approach, between two different datasets, in order to obtain results able to simulate a real environment.

Race classification involves some complications related to the soft appearance threshold that usually involves race groups. Lu and Jain [24] use LDA to classify Asian vs non Asian people, Toderici et al. [39] use 3D model with the purpose to improve race classification.

Aging estimation is one of the most investigated and not trivial topic in demographic classification. Indeed people with the same calendar age could exhibit highly different biological age because of an harder or relaxed lifestyle. AAM is a widely used techniques for age group classification used for instance by Liu et al. [22]. Doung et al. combine Active Appearance Models (AAMs) technique and LBP local facial features in combination, while Ylioinas et al. exploit the variant of LBP features to encode facial micro-patterns [42,43].

Moreover some works are interested in the estimation of the whole set of demographic information. Hadid and Pietikäinen [16] exploits spatio-temporal information from video sequence and analyze the correlation between the face images through manifold learning. Klare et al. [19] study the influence of demographic appearance in face recognition.

We choose to compare three different features as Local Binary Pattern (LBP), Histogram of Oriented Gradient (HOG) and Weber Local Descriptor (WLD) and two subspace reduction approaches as Principal Component Analysis (PCA) and Fisher Linear Discriminant Analysis (LDA) for all gender race and age group classification, with the aim to discuss the trade-off between the use of unique or more methods and the computational effort.

Moreover we test the possibility to improve age accuracy using an age with race knowledge classifier.

The article is organized in 5 sections. After the introduction we expose the proposed pipeline for face detection and classification in Sect. 2. The section is also aimed to give an overview of the used methods. Section 3 presents a wide description of employed data-set and their configuration in experiments. Section 4 describes the experimental configuration and presents and discusses the results. Conclusions are finally discussed in Sect. 5.

## 2 Demographic Face Classification Pipeline

Face classification from generic video sequences involves a complex pipeline of operations that could involve different operative blocks. Our choice is to use a model as in Fig. 1. First step is to look for faces in the actual image. When the faces in the scene have been detected, the classification can be done.

Gender, race, or age classification needs the extraction and analysis of specific features capable to discriminate classes of interest as much as possible. Anyway, the extraction of features needs to be consistent among faces. For this reason a geometrical normalization of the face image is necessary. Normalization performs

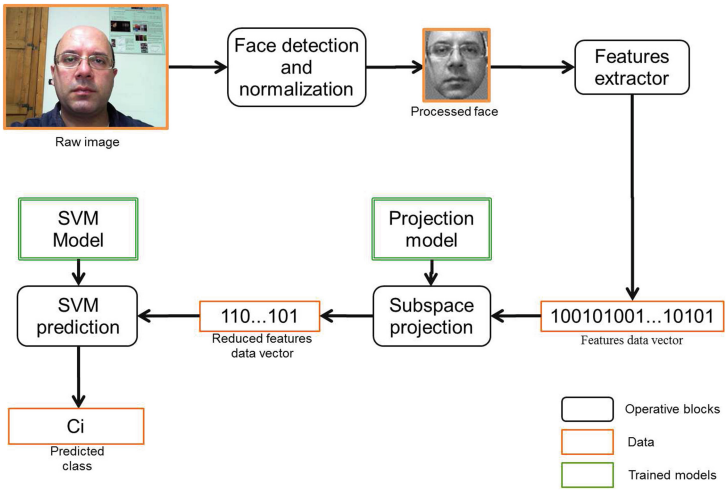



**Fig. 1.** Demographic face classification pipeline: the raw image is processed in order to obtain a reliable prediction for gender/race/age of the subject under test.

geometrical transformation on the face image in such a manner to obtain a standard configuration relative to eye, mouth and nose position. This is necessary to ensure that the spatial features extraction is coherently referred to the same portion of face image among different faces. Features extraction is than applied to the normalized face.

The set of features is then projected, by means of a well trained subspace operator (PCA, LDA), in a reduced and more discriminating subspace, performing a considerable reduction of components in the features vector. This kind of reduction is a key factor to obtain a decision algorithm (we employed a SVM approach) computationally efficient both in terms of training and prediction.

The last step is the prediction by means of the support vector machines (SVM) model. SVM returns a class prediction based on a model adequately trained.

Our purpose is to find a pair of "features vector"/"subspace projection" capable to exhibit a reasonable trade-off for all the three needed demographic information. A unique type of features means an high memory and computational save really useful in systems as embedded platforms for digital signage where the computational resources could be limited.

### 2.1 Face Detection and Normalization

Generic environment of interest for face classification usually involves scenario with one or more faces. The preprocessing consists in detection of the faces in the scene and the consecutive normalization (Fig. 2). This problem has been well treated by Castrillón et al. that in [6] develop ENCARA2 library in order to perform a face detection and normalization process frame by frames.


**Fig. 2.** Face detection and normalization: A Viola-Jones face detector detect the faces present in an image. When the face is well located in the space an eye detection is used to rotate and scale the face in such a manner that all face present the same scale and orientation

They mainly implement the well known Viola-Jones [41] face detector to detect faces in the current frame. Successively an eye detection is performed to locate the eye pairs in the image and to rotate and scale the face with the aim of obtaining standard face images with eyes pair located in the same position.

### 2.2 Features Extraction

The classification process needs a suitable features vector able to characterize, as much as possible, the main aspects for gender, race and age. We chose to work with three light and well known descriptor designed in such a way to recognize different informative components of the analyzed image.

**LBP.** Local Binary Pattern (LBP) is a local feature widely utilized for texture description in pattern recognition. The original LBP assign a label to each pixel. Each pixel is used as a threshold and compared with its neighbor. If the neighbor is higher it takes the value 1 otherwise it takes the value 0. Finally the thresholded neighbor pixel values are concatenated and considered as a binary number that becomes the label for the central pixel. A graphical representation is showed in Fig. 3 (a). In formula

$$LBP_{P,R}(x_c) = \sum_{p=0}^{P-1} u(x_p - x_c)2^p \tag{1}$$

where $u(y)$ is the step function, $x_p$ the neighbor pixels, $x_c$ the central pixel $P$ the number of neighbors and $R$ the radius. To account for the spatial information the image is then divided in sub-regions. LBP is applied to each sub-region and an histogram of L bins is generated from the pixel labels. Then the histograms of different regions are concatenated in a single higher dimensional histogram as represented in Fig. 3 (b).

**HOG.** Local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. HOG is a well known feature descriptor based on the accumulation of gradient directions over the pixel of a small spatial region referred as a "cell", and in the consequent construction of a 1D histogram. Even though HOG has many precursors, it has been used in its mature form in Scale Invariant Features Transformation [23] and

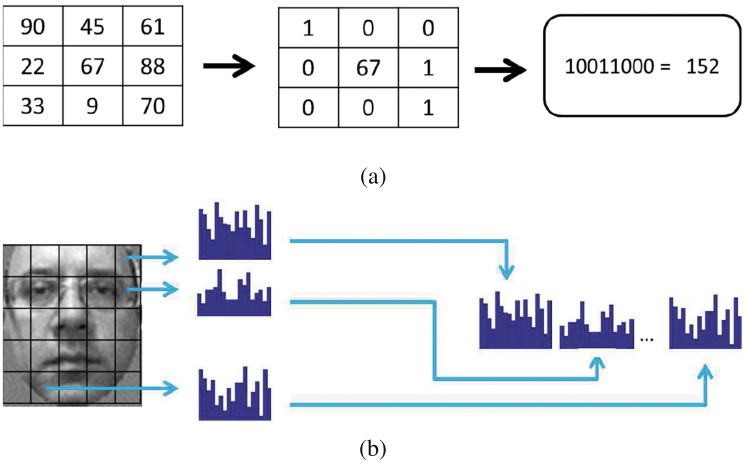



**Fig. 3.** LBP labeling procedure (a) and spatial LBP histogram concatenation (b)

widely analyzed in human detection by Dalal and Triggs [12]. This method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. Let L be the image to analyze. The image is divided in cells (Fig. 4 (a)) of size $N \times N$ pixels and the orientation $\theta$ of each pixel $x = (x_x, x_y)$ is computed (Fig. 4 (b)) by means of the following rule:

$$\theta(x) = \tan^{-1} \frac{L(x_x, x_y + 1) - L(x_x, x_y - 1)}{L(x_x + 1, x_y) - L(x_x - 1, x_y)} \tag{2}$$

The orientations are accumulated in an histogram of a predetermined number of bins (Fig. 4 (c–d)). Finally histograms of each cell are concatenated in a single spatial HOG histogram (Fig. 4 (e)). In order to achieve a better invariance to disturbs, it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram energy over larger spatial regions, named blocks, and using the results to normalize all of the cells in the block. The normalized descriptor blocks will represent the HOG descriptors.

**SWLD.** Weber Local Descriptor (WLD) [9] is a robust and powerful descriptor inspired to Weber's law. It is based on the fact that the human perception of a pattern depends not only on the amount of change in intensity of the stimulus but also on the original stimulus intensity. The proposed descriptor consist of two components: differential excitation (DE) and gradient orientation (OR).

Differential excitation allows to detect the local salient pattern by means of a ratio between the relative intensity difference of the current pixel against its neighbors and the intensity of the current pixel. The computational approach is very similar to LBP one (for a deeply description we remand to the main paper [9]). Moreover the OR of the single pixel is considered.

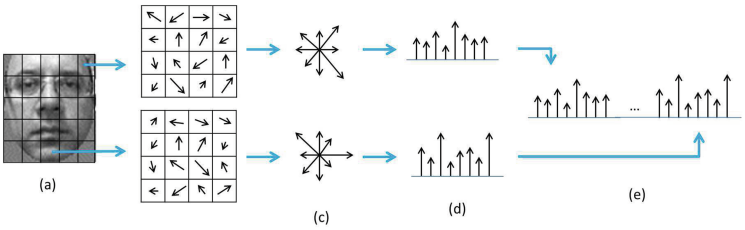



**Fig. 4.** HOG extraction features representation. Image is divided in cells. Each cell is done of $N \times N$ pixel. We compute the orientation of all pixel and construct the histogram of orientation of the cell. Finally all orientation histogram are concatenated to construct the final features vector.

When DE and OR are computed for each pixel in the image we construct a 2D histogram of $T$ columns and $M \times S$ rows where $T$ is the number of orientations and $M \times S$ the number of bins for the DE quantization with the meaning of [9]. Figure 5 shows how 2D histogram is mixed in such a way to obtain a 1D histogram more suitable for successive operation.

Anyway our purpose is to account also for spatial information. With this aim we choose to use the Spatial WLD approach (SWLD) [40] that, in the same way of spatial LBP, splits the image in sub-regions and computes an histogram for each of them. Finally, the histograms are concatenated in an ordered way.

### 2.3 Features Subspace Projection

The number of used features for face description is highly influent in computational complexity and accuracy of classification. Indeed a reduced number of features allows SVM to use easier functions and to perform better division of cluster. Anyway the reduction of original features space is a non trivial step.

**PCA.** Principal component analysis (PCA) is a widely used approach for subspace reduction. It chooses a dimensionality reducing linear projection that maximizes the scatter of all projected samples. Simply speaking the more informative subspace directions are selected for the subspace reduction. The number of components should be previously selected and, in the specific case of face recognition and analysis, it could be large.

**LDA.** Linear Discriminant Analysis (LDA) [3] is an alternative subprojection approach. In LDA, within-class and between-class scatters are used to formulate criteria for class separability. The optimizing criterion in LDA is the ratio of between-class scatter to the within-class scatter. The solution obtained by maximizing this criterion defines the axes of the transformed space. Moreover, in LDA analysis the number of non-zero generalized eigenvalue, and so the upper-bound in eigenvectors numbers, is $c - 1$ where $c$ represent the number of class.

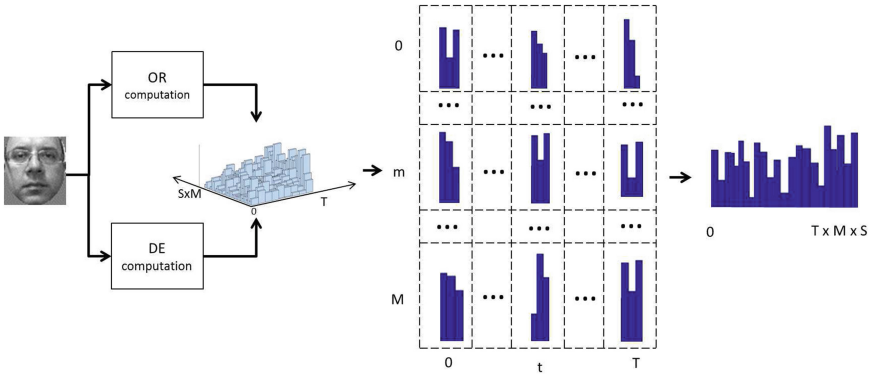



**Fig. 5.** WLD histogram construction process: the algorithm compute the DE and OR values for each pixel and construct the 2D histogram. The 2D histogram is splitted in a M by T matrix where each element is an histogram of S bins. Finally the whole 1D histogram is composed by the concatenation of the previous matrix rows.

Belhumeur et al. [3] show the robustness of this approach for face recognition and show how LDA can be more discriminative in some situations. We would like to remark that the upper bound of eigenvectors for LDA projection approach is a key factor also in terms of computational complexity that makes the workload of SVM lighter.

### 2.4 SVM Prediction

After data projection, in the proposed approach gender, race and age classification are performed by using Support Vector Machines (SVM). In particular three different SVMs are used: the gender classification problem uses the classical two-class SVM approach in order to discriminate between male and female human gender, whereas race and age group classification problems are solved by means of multi-class SVM.

SVM is a discriminative classifier defined by a separating hyperplane. Given a set of labeled training data (supervised learning), the algorithm computes an optimal hyperplane (the trained model) which categorizes new examples in the right class. In particular the $C$-support vector classification ($C$-SVC) learning task implemented in the well-known LIBSVM [8] has been used. Given training vectors $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, \cdots, l$, in two classes, and an label vector $\boldsymbol{y} \in \mathbb{R}^l$ such that $y_i \in \{1, -1\}$, $C$-SVC [4,10] solves the following primal optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \cdots, l, \end{aligned}$$

where $\phi(\boldsymbol{x}_i)$ maps $\boldsymbol{x}_i$ into a higher-dimensional space and $C > 0$ is the regularization parameter. Due to the possible high dimensionality of the vector variable $\boldsymbol{w}$, usually the following dual problem is solved:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \boldsymbol{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & \boldsymbol{y}^T\boldsymbol{\alpha} = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, \cdots, l, \end{aligned}$$

where $\boldsymbol{e} = [1, \cdots, 1]^T$ is the vector of all ones, $Q$ is an $l \times l$ positive semidefinite matrix, $Q_{ij} y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $K(\boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ is the kernel function. After the dual problem is solved, the next step is to compute the optimal $w$ as:

$$\boldsymbol{w} = \sum_{i=1}^{l} y_i \alpha_i \phi(\boldsymbol{x}_i) \tag{3}$$

Finally the decision function is

$$\mathtt{sgn}(\boldsymbol{w}^T \phi(\boldsymbol{x}) + b) = \mathtt{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \tag{4}$$

Such an approach is suitable only for the two classes gender problems. For race and age group classification a multi-class approach has been used. It can be treated through the "one-against-one" [20]. Let $k$ be the number of classes, then $k(k-1)/2$ classifiers are constructed where each one trains data from two classes. The final prediction is returned by a voting system among all the classifiers. Many other methods are available for multi-class SVM classification, anyway in [18] a detailed comparison is given with the conclusion that "one-against-one" is a competitive approach.

## 3 Dataset Description

The choice of the dataset is a non trascurable issue in the performance estimation of demographic recognizing algorithms. Indeed classes distribution and number of examples can highly influence the generalization of the results. With the end to mitigate the unbalancing among classes we used a fused dataset made-up by two of the most representative datasets in face classification problems: Morph [1] and Feret [30]. Both datasets consist of face images of people of different gender, ethnicity and age. Morph presents only frontal faces, on the other hand Feret is made up by feces in different poses from which we considered just the frontal and partially rotated faces. Both datasets are provided with a complete CVS annotation file with gender, race and other information.

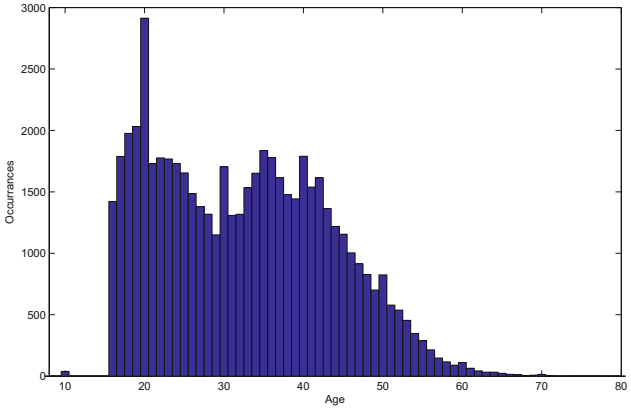



**Fig. 6.** Age histogram of the dataset.

Anyway, face detection is not always full reliable and a reasonable number of misses in face detection and normalization is reasonable. Due to this issue, down-line of the face detection and normalization, the dataset presents 55915 male subjects and 9246 female subjects. It is also partitioned in 41334 Black people, 11843 White people, 601 Asian people and finally 1889 Hispanic. The age distribution is represented in Fig. 6.

Moreover, the dataset has been treated in order to obtain balanced subsets for each gender and race problems (regarding the age problem the dataset used didn't permit to extract an adequate balanced subset). To this end an upper-bound (lower than the less numerous class) is selected. Then the new subset is created filling each class randomly selecting entries from the full dataset until the upper-bound is reached. In this way additional balanced datasets have been built. More specifically the gender subset counts 1701 entries for class and 500 element for the race ones.

## 4 Experimental Results

We focus our experimental set-up to find the best trade-off between demographic estimation accuracy and computational complexity. We analyze performances of the three methods presented in combination with two possible subspace approaches.

We test the three most important demographic aspects in estimation application of commercial interest: Gender, Race and Age. More specifically race has been divided in 4 possible classes: White, Black, Asian, Hispanic and age in three classes: Young (0 to 29 years old), Middle (30 to 50 years old), Mature (51 and more years old).

We performed face detection and normalization over each facial image in order to obtain a $59 \times 65$ grey scale facial images. Successively we extract the features vector (features of the method under test).

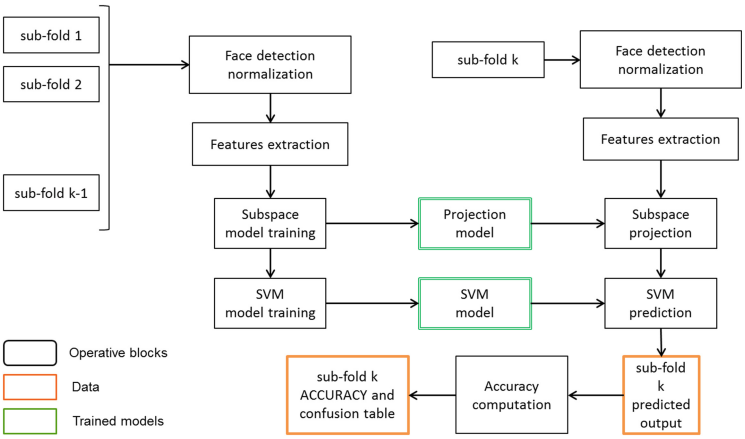



**Fig. 7.** Test procedure for accuracy estimation: the procedure is done $k$ times in order to obtain the best estimation of total accuracy and confusion table.

The accuracy estimation of the proposed prediction algorithms has been performed following the procedure showed in Fig. 7. It consists of two steps: a model estimation and a prediction estimation. The whole dataset has been randomly split in $k$ sub-folds. For each of the $k$ validation steps, $k-1$ sub-fold for the training and 1 sub-fold for the prediction/validation process have been used. Face detection and normalization have been then performed on each image of the selected $k-1$ training sub-fold and then the *data vectors* have been extracted. The set of *data vectors* has been then used to train, in sequence, the feature reduction algorithm and the SVMs. Finally the one-out fold has been finally tested by using the available models. The process is repeated over each of the five fold configuration and the accuracy results are averaged.

Moreover, each test has been also performed using an additional data scaling transformation. Anyway, in order to simplify the results presentation, just the better results, either scaled or non scaled version, have been showed.

LBP has been spatially processed over a $5 \times 5$ grid of the image with 8 neighbors and 1 pixel radius. We employed an online implementation of LBP[1].

HOG space unit has been selected as a $4 \times 4$ pixel size sub image section and 9 orientation/bins. The *VLFeat library* is used for HOG operator[2].

SWLD (developed by the authors following [40]) has been computed with value of T, M, S that are respectively of 8, 4, 4 over $4 \times 4$ grid of the image.

Moreover, in our experience, a number of 100 component for the PCA was taken into account in order to preserve the 95 % of the total variance of data.

Finally we adopt a SVM prediction model with radial basis function and standard parameter of LIBSVM library [8]. We used, for all the three demographic estimations, a radial basis function (RBF) that, in the opinion of the authors of [17] as well as in our experience, seems to be the most reasonable choice.

[1] www.bytefish.de/blog/local_binary_patterns/
[2] www.vlfeat.org

Usually a grid search for penalty parameter $C$ and the other RBF parameters could be desirable. Anyway, in our tests any significant difference in the results arises as the parameters change. More specifically, we set $C = 1$ and $\gamma = 1/N_f$ where $N_f$ is the number of features.

We have chosen to present all results in terms of accuracy confusion tables in order to manage possible blunders introduced by the different numerosity among classes. More specifically each row of the table is normalized and expressed in terms of percentage. With similar purpose the total accuracy (TA) has been expressed as the mean of the diagonal entries of the obtained confusion matrix. Indeed the computation of accuracy as the rate between the True positive and the total number of tested entries could be an unreasonable result if the numerosity among classes is not comparable (Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18).

An overview, of the obtained results, shows that the most performing descriptor is the HOG one. Indeed, among gender, race and age problems it seems able to guarantee, in association with PCA or LDA reduction approaches, the most stable and accurate predictions. Only for gender estimation, SWLD + LDA (unbalanced dataset) outperforms HOG + PCA and HOG + LDA (balanced subset) with an accuracy of 90.0 % in the unbalanced dataset configuration. However it has to be noticed that HOG + PCA's true positive rates are better distributed among Male and Female classes with an total accuracy of only 0.3 % less than SWLD + LDA's one. Similar considerations are true for the race problem where HOG + PCA (balanced subset), HOG + LDA (unbalanced dataset) and SWLD + LDA (unbalanced dataset) present similar performances. Another time the first one guarantee a better distribution of true positive rates among races against the HOG + LDA and SLWD + LDA cases that show excellent performance for White and Black classes and unreliable performances for the Asian and Hispanic ones.

Probably, due to the fact that HOG looks to orientation appearance that is more discriminant then the texture analysis performed by the LBP, the accuracy of the LBP based approaches decreases considerably and exhibits value not still reliable. On the other hand SWLD with LDA configuration, reaches results similar to the HOG features. SWLD accounts for both texture and orientation appearance and the good performances in LDA configuration seems reasonable. To the opposite SWLD + PCA exhibits worse performances probably due to the inability of PCA to find the best subspace reduction for the specific features/class classification problem (Table 19).

The analysis of the results of the balanced and unbalanced datasets opens another discussion. Indeed it is clear that the unbalanced dataset guarantees high accuracy results for the most numerous classes and worse results for classes with few entries. On the other hand, using the balanced dataset (with entries for class reduced to the less numerous class) the accuracy over the previously more numerous classes decrees and the accuracy over the previously poor classes increases with an average accuracy that usually outperforms the previous one.

**Gender confusion tables:** each table presents the results for the each specific descriptor/projection pair for both balanced and unbalanced data-set configuration. $M_T$: Male true; $F_T$: Female true; $M_P$: Male prediction; $F_P$: Female prediction; $M_P^B$: Male prediction using balanced data-set; $F_P^B$: Female prediction using balanced data-set; TA: Total accuracy; $\text{TA}^B$: Total accuracy using balanced data-set

**Table 1.** LBP + PCA

| | $M_T$ | $F_T$ |
|---|---|---|
| $M_P$ | 99.3 % | 0.7 % |
| $F_P$ | 58.4 % | 41.5 % |
| $M_P^B$ | 84.7 % | 15.3 % |
| $F_P^B$ | 25.1 % | 74.9 % |
| TA | 70.1 % | |
| $\text{TA}^B$ | 79.8 % | |

**Table 2.** LBP + LDA

| | $M_T$ | $F_T$ |
|---|---|---|
| $M_P$ | 100 % | 0 % |
| $F_P$ | 100 % | 0 % |
| $M_P^B$ | 82.9 % | 17.1 % |
| $F_P^B$ | 15.3 % | 84.7 % |
| TA | 50 % | |
| $\text{TA}^B$ | 83.8 % | |

**Table 3.** HOG + PCA

| | $M_T$ | $F_T$ |
|---|---|---|
| $M_P$ | 97.8 % | 2.2 % |
| $F_P$ | 25.8 % | 74.2 % |
| $M_P^B$ | 87.3 % | 12.7 % |
| $F_P^B$ | 7.8 % | 92.2 % |
| TA | 86 % | |
| $\text{TA}^B$ | 89.7 % | |

**Table 4.** HOG + LDA

| | $M_T$ | $F_T$ |
|---|---|---|
| $M_P$ | 98.7 % | 1.3 % |
| $F_P$ | 21.5 % | 78.5 % |
| $M_P^B$ | 82.1 % | 17.9 % |
| $F_P^B$ | 21.2 % | 78.8 % |
| TA | 88.6 % | |
| $\text{TA}^B$ | 80.5 % | |

**Table 5.** SWLD + PCA

| | $M_T$ | $F_T$ |
|---|---|---|
| $M_P$ | 100 % | 0 % |
| $F_P$ | 100 % | 0 % |
| $M_P^B$ | 85.9 % | 14.1 % |
| $F_P^B$ | 22.1 % | 77.9 % |
| TA | 50 % | |
| $\text{TA}^B$ | 81.9 % | |

**Table 6.** SWLD + LDA

| | $M_T$ | $F_T$ |
|---|---|---|
| $M_P$ | 97.7 % | 2.3 % |
| $F_P$ | 17.7 % | 82.3 % |
| $M_P^B$ | 74.6 % | 25.4 % |
| $F_P^B$ | 24.5 % | 75.5 % |
| TA | 90 % | |
| $\text{TA}^B$ | 75.5 % | |

**Race confusion tables:** each table presents the results for the each specific descriptor/projection pair for both balanced and unbalanced data-set configuration. $W_T$: White true; $B_T$: Black true; $A_T$: Asian true; $H_T$: Hispanic true; $W_P$: White prediction; $B_P$: Black prediction; $A_P$: Asian prediction; $H_P$: Hispanic prediction; $W_P^B$: White prediction using balanced data-set; $B_P^B$: Black prediction using balanced data-set; $A_P^B$: Asian prediction using balanced data-set; $H_P^B$: Hispanic prediction using balanced data-set; TA: Total accuracy; $\text{TA}^B$: Total accuracy using balanced data-set

**Table 7.** LBP + PCA

| | $W_T$ | $B_T$ | $A_T$ | $H_T$ |
|---|---|---|---|---|
| $W_T$ | 93.2 % | 6.8 % | 0 % | 0 % |
| $B_T$ | 1.6 % | 98.4 % | 0 % | 0 % |
| $A_T$ | 83.2 % | 16.8 % | 0 % | 0 % |
| $H_T$ | 58.5 % | 41.5 % | 0 % | 0 % |
| $W_T^B$ | 49 % | 14.8 % | 12.8 % | 23.4 % |
| $B_T^B$ | 1.2 % | 91.6 % | 0.2 % | 7 % |
| $A_T^B$ | 3.6 % | 8.2 % | 73.4 % | 14.8 % |
| $H_T^B$ | 14.6 % | 23.2 % | 6.2 % | 56 % |
| TA | 47.9 % | | | |
| $\text{TA}^B$ | 67.5 % | | | |

**Table 8.** LBP + LDA

| | $W_T$ | $B_T$ | $A_T$ | $H_T$ |
|---|---|---|---|---|
| $W_P$ | 0 % | 100 % | 0 % | 0 % |
| $B_P$ | 0 % | 100 % | 0 % | 0 % |
| $A_P$ | 0 % | 100 % | 0 % | 0 % |
| $H_P$ | 0 % | 100 % | 0 % | 0 % |
| $W_P^B$ | 70.2 % | 3.4 % | 7.4 % | 19 % |
| $B_P^B$ | 4.6 % | 84.8 % | 2.2 % | 8.4 |
| $A_P^B$ | 7.6 % | 3 % | 78 % | 11.4 % |
| $H_P^B$ | 20.2 % | 6 % | 9.8 % | 64 % |
| TA | 25 % | | | |
| $\text{TA}^B$ | 74.2 % | | | |

**Table 9.** HOG + PCA

| | $W_T$ | $B_T$ | $A_T$ | $H_T$ |
|---|---|---|---|---|
| $W_P$ | 92.9 % | 7.1 % | 0 % | 0 % |
| $B_P$ | 1.7 % | 98.3 % | 0 % | 0 % |
| $A_P$ | 58.4 % | 40.8 % | 0 % | 0.8 % |
| $H_P$ | 67.0 % | 27.3 % | 0 % | 5.7 % |
| $W_P^B$ | 68.2 % | 2.6 % | 12.2 % | 17 % |
| $B_P^B$ | 2.4 % | 89 % | 2 % | 6.6 % |
| $A_P^B$ | 2 % | 3.4 % | 75 % | 19.6 % |
| $H_P^B$ | 15.2 % | 5 % | 7.4 % | 72.4 % |
| TA | 49.2 % | | | |
| $\text{TA}^B$ | 76.1 % | | | |

**Table 10.** HOG + LDA

| | $W_T$ | $B_T$ | $A_T$ | $H_T$ |
|---|---|---|---|---|
| $W_P$ | 94.3 % | 3.4 % | 0.4 % | 1.9 % |
| $B_P$ | 1 % | 98.5 % | 0.1 % | 0.4 % |
| $A_P$ | 19.3 % | 11.3 % | 59.4 % | 10 % |
| $H_P$ | 30.1 % | 16.2 % | 2.2 % | 51.5 % |
| $W_P^B$ | 44.6 % | 10 % | 16.6 % | 28.8 % |
| $B_P^B$ | 12.6 % | 58.8 % | 13.2 % | 15.4 % |
| $A_P^B$ | 14.6 % | 8.2 % | 57 % | 20.2 % |
| $H_P^B$ | 25.8 % | 13 % | 15.2 % | 46 % |
| TA | 75.9 % | | | |
| $\text{TA}^B$ | 51.6 % | | | |

**Table 11.** SWLD + PCA

| | $W_T$ | $B_T$ | $A_T$ | $H_T$ |
|---|---|---|---|---|
| $W_P$ | 88.7 % | 11.3 % | 0 % | 0 % |
| $B_P$ | 1.6 % | 98.4 % | 0 % | 0 % |
| $A_P$ | 85.5 % | 14.5 % | 0 % | 0 % |
| $H_P$ | 54.5 % | 45.5 % | 0 % | 0 % |
| $W_P^B$ | 47.8 % | 9.2 % | 12.8 % | 30.2 % |
| $B_P^B$ | 1.2 % | 93.6 % | 0.2 % | 5 % |
| $A_P^B$ | 3.8 % | 7.6 % | 73.4 % | 15.2 % |
| $H_P^B$ | 12 % | 23.2 % | 6.2 % | 58.6 % |
| TA | 46.77 % | | | |
| $\text{TA}^B$ | 68.3 % | | | |

**Table 12.** SWLD + LDA

| | $W_T$ | $B_T$ | $A_T$ | $H_T$ |
|---|---|---|---|---|
| $W_P$ | 94.6 % | 3.3 % | 0.5 % | 1.6 % |
| $B_P$ | 1.2 % | 98.5 % | 0 % | 0.3 % |
| $A_P$ | 23.5 % | 10.8 % | 58.9 % | 8.8 % |
| $H_P$ | 37.7 % | 19.9 % | 2.4 % | 39.9 % |
| $W_P^B$ | 51.2 % | 9.8 % | 14 % | 25 % |
| $B_P^B$ | 12.6 % | 58.8 % | 9.6 % | 19 % |
| $A_P^B$ | 15 % | 6.2 % | 61.8 % | 17 % |
| $H_P^B$ | 22.4 % | 17.6 % | 16.4 % | 43.9 % |
| TA | 72.5 % | | | |
| $\text{TA}^B$ | 53.8 % | | | |

**Age confusion tables:** each table presents the results for each the specific descriptor/projection pair. $Y_T$: true class for people aged under 26 years; $M_T$: true class for people aged between 27 and 49 years; $O_T$: true class for people older than 50 years; $Y_P$: predicted class for people aged under 26 years; $M_P$: predicted class for people aged between 27 and 49 years; $O_P$: predicted class for people older than 50 years; TA: total accuracy

**Table 13.** LBP + PCA

| | $Y_T$ | $M_T$ | $O_T$ |
|---|---|---|---|
| $Y_P$ | 44.7 % | 55.3 % | 0 % |
| $M_P$ | 7.9 % | 92.1 % | 0 % |
| $O_P$ | 0.7 % | 99.3 % | 0 % |
| TA | 45.6 % | | |

**Table 14.** LBP + LDA

| | $Y_T$ | $M_T$ | $O_T$ |
|---|---|---|---|
| $Y_P$ | 0 % | 100 % | 0 % |
| $M_P$ | 0 % | 100 % | 0 % |
| $O_P$ | 0 % | 100 % | 0 % |
| TA | 33.3 % | | |

**Table 15.** HOG + PCA

| | $Y_T$ | $M_T$ | $O_T$ |
|---|---|---|---|
| $Y_P$ | 67.7 % | 32.3 % | 0 % |
| $M_P$ | 13.1 % | 86.9 % | 0 % |
| $O_P$ | 0.7 % | 99.3 % | 0 % |
| TA | 51.5 % | | |

**Table 16.** HOG + LDA

| | $Y_T$ | $M_T$ | $O_T$ |
|---|---|---|---|
| $Y_P$ | 72.3 % | 27.7 % | 0 % |
| $M_P$ | 11.5 % | 86.7 % | 1.8 % |
| $O_P$ | 0.4 % | 66 % | 33.6 % |
| TA | 64.2 % | | |

**Table 17.** SWLD + PCA

| | $Y_T$ | $M_T$ | $O_T$ |
|---|---|---|---|
| $Y_T P$ | 25.8 % | 74.2 % | 0 % |
| $M_P$ | 4.5 % | 95.5 % | 0 % |
| $O_P$ | 0.3 % | 99.7 % | 0 % |
| TA | 40.4 % | | |

**Table 18.** SWLD + LDA

| | $Y_T$ | $M_T$ | $O_T$ |
|---|---|---|---|
| $Y_P$ | 69.4 % | 30.6 % | 0 % |
| $M_P$ | 13 % | 85.2 % | 1.8 % |
| $O_P$ | 0.6 % | 67.5 % | 31.9 % |
| TA | 62.1 % | | |

Due to variety of possible uses illustrated in the introduction and through the paper, it results clear that the computational complexity and memory load is a key aspect in these kind of field. First of all we analyze the average "elaboration time" for LBP, HOG and SWLD over 1000 iteration. It results that:

- LBP = 0.376 ms
- HOG = 0.392 ms
- SWLD = 1.150 ms

where the time has been computed over a Xeon E5 (12 core) with 32 Gb RAM. LBP and HOG results faster than SWLD whose implementation could be, in our opinion, improved.

Regarding data reduction approaches, PCA reduces the size of features vector considerably but a number of 100 components are necessary to keep desirable accuracy performance. This limitation reduces the possible advantage introduced by subspace projection. On the other hand LDA allows a considerable reduction in memory occupation for both the projection model and the features vector and consequently a reduction of the number of operations in subspace projection. This aspect becomes critical in embedded systems (widely used in mobile and real time applications) where the workload and memory occupation reduction could allow the grow of image size for analysis. Results confirm that the use of LDA reduces considerably (about 1 magnitude) the elaboration time for both the projection step and SVM prediction.

Finally the SVM prediction stage shows elaboration time that is quite similar among the different classification problems and used descriptors. This is reasonable because, independently of the class problem, the number of components for the PCA is the same and leads to an elapsed time of 0.01 ms. The computational time for SVM prediction, down-line the LDA projection, is unappreciative due to the few number of elements in the vector (3 at the most).

Let us consider our normalized $59 \times 65$ pixels image. It consists, for LBP + PCA approach, of a 6400 features vector reduced to 100 components through a $6400 \times 100$ size projection matrix and 640000 operations. If we would double the image size, the starting features vector, the projection matrix and the number of operations necessary to the projection, would grow of 4 times the original size leading to a projection matrix of $25600 \times 100$ and a number of operations of 2560000. It is clear that the necessity to work with higher resolution image could be rapidly become an intractable problem, at least for low computational power embedded system. Opposite the 97 % to 99 % percentage reduction of projection matrix dimensions and operations needed for projection introduced by the use of LDA allows to work with an image resolution more suitable for analysis of texture aspects such as wrinkles and beard, fundamentals for age estimation. Table 20 summarizes features vector and number of operations necessary for subspace projection in each possible configuration illustrated through the paper.

**Table 19.** Elaboration time (sec) for the projection step (averaged over 500 trials)

| | LBP | | HOG | | SWLD | |
|---|---|---|---|---|---|---|
| | PCA | LDA | PCA | LDA | PCA | LDA |
| Gender | 0.3125 | 0.01 | 0.0925 | 0.0025 | 0.115 | 0.0025 |
| Race | 0.32 | 0.0225 | 0.095 | 0.0075 | 0.095 | 0.0075 |
| Age | 0.36 | 0.02 | 0.106 | 0.006 | 0.11 | 0.006 |

**Table 20.** The tables present the size of original and subspace projected features vector (a) and the number of operation for subspace reduction (b). (LDA-G = LDA with gender, LDA-R = LDA with race, LDA-A = LDA with age)

(a)

| | LBP | HOG | SWLD |
|---|---|---|---|
| Base | 6400 | 2016 | 2048 |
| PCA | | 100 | |
| LDA-G | | 1 | |
| LDA-R | | 3 | |
| LDA-A | | 2 | |

(b)

| | LBP | HOG | SWLD |
|---|---|---|---|
| PCA | 640,000 | 201,600 | 204,800 |
| LDA-G | 6,400 | 2,016 | 2,048 |
| LDA-R | 19,200 | 6,048 | 6,144 |
| LDA-A | 12,600 | 4,032 | 4,096 |

## 5 Conclusions

We analyzed different combinations of descriptors and subspace projections for gender, race and age classification problems. Results show that HOG and SWLD descriptors are the most robust and performant among different classification problems. Anyway, the choice of PCA or LDA as subspace projections, as well as the balanced or unbalanced sub datasets for training is highly influent to get the best performances from the descriptor. The analysis of the results shows the importance of balancing among classes in order to obtain constant true positive rate for each class. Moreover the dominance of a single descriptor is really useful because allows the computation of a single features vector to be processed in different ways for all the demographic aspects. This is a key point for systems where the computational effort and memory are a precious resource both in terms of energy save and system capacity limits.

## References

1. Morph-noncommercial face dataset. http://www.faceaginggroup.com/morph/
2. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recogn. Lett. **36**(0), 228–234 (2014). http://www.sciencedirect.com/science/article/pii/S0167865513001864
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
4. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. COLT '92, pp. 144–152. ACM, New York (1992). http://doi.acm.org/10.1145/130385.130401
5. Brunelli, R., Poggio, T.: HyberBF networks for gender classification (1995)

6. Castrillón, M., Déniz, O., Guerra, C., Hernández, M.: ENCARA2: real-time detection of multiple faces at different resolutions in video streams. J. Vis. Commun. Image Represent. **18**(2), 130–140 (2007)
7. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Improving gender classification accuracy in the wild. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) CIARP 2013, Part II. LNCS, vol. 8259, pp. 270–277. Springer, Heidelberg (2013). http://dx.doi.org/10.1007/978-3-642-41827-3_34
8. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/~cjlin/libsvm
9. Chen, J., Shan, S., He, C., Zhao, G., Pietikäinen, M., Chen, X., Gao, W.: WLD: a robust local image descriptor. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1705–1720 (2010)
10. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). http://dx.doi.org/10.1023/A%3A1022627411411
11. Dago-Casas, P., Gonzalez-Jimenez, D., Yu, L.L., Alba-Castro, J.: Single- and cross-database benchmarks for gender classification under unconstrained settings. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2152–2159, Nov 2011
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, pp. 886–893, June 2005
13. Davies, N., Langheinrich, M., Jose, R., Schmidt, A.: Open display networks: a communications medium for the 21st century. Computer **45**(5), 58–64 (2012)
14. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
15. Abdi, H., Valentin, D., Edelman, B., O'Toole, A.J.: More about the difference between men and women: evidence from linear neural networks and the principal-component approach. Neural Comput. **7**(6), 1160–1164 (1995)
16. Hadid, A., Pietikäinen, M.: Demographic classification from face videos using manifold learning. Neurocomputing **100**(0), 197–205 (2013). http://www.sciencedirect.com/science/article/pii/S0925231212003906, Special issue: Behaviours in video
17. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
18. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. **13**(2), 415–425 (2002)
19. Klare, B., Burge, M., Klontz, J., Vorder Bruegge, R., Jain, A.: Face recognition performance: role of demographic information. IEEE Trans. Inf. Forens. Secur. **7**(6), 1789–1801 (2012)
20. Knerr, S., Personnaz, L., Dreyfus, G.: Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Soulié, F., Hérault, J. (eds.) Neurocomputing. NATO ASI Series, vol. 68, pp. 41–50. Springer, Heidelberg (1990). http://dx.doi.org/10.1007/978-3-642-76153-9_5
21. Krumm, J.: Ubiquitous advertising: the killer application for the 21st century. IEEE Perv. Comput. **10**(1), 66–73 (2011)
22. Liu, L., Liu, J., Cheng, J.: Age-group classification of facial images. In: 2012 11th International Conference on Machine Learning and Applications (ICMLA), vol. 1, pp. 693–696, Dec 2012
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

24. Lu, X., Jain, A.K.: Ethnicity identification from face images. In: Proceedings of the SPIE Defense and Security Symposium, Orlando, FL, pp. 165–170, Apr 2004
25. Lyons, M.J., Budynek, J., Plante, A., Akamatsu, S.: Classifying facial attributes using a 2-d Gabor wavelet representation and discriminant analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 202–207 (2000)
26. Michelis, D., Müller, J.: The audience funnel: observations of gesture based interaction with multiple large displays in a city center. Int. J. Hum. Comput. Inter. **27**(6), 562–579 (2011)
27. Mäkinen, E., Raisamo, R.: An experimental comparison of gender classification methods. Pattern Recogn. Lett. **29**(10), 1544–1556 (2008). http://www.sciencedirect.com/science/article/pii/S0167865508001116
28. Müller, J., Wilmsmann, D., Exeler, J., Buzeck, M., Schmidt, A., Jay, T., Krüger, A.: Display blindness: the effect of expectations on attention towards digital signage. In: Tokuda, H., Beigl, M., Friday, A., Brush, A.J.B., Tobe, Y. (eds.) Pervasive 2009. LNCS, vol. 5538, pp. 1–8. Springer, Heidelberg (2009)
29. Müller, J., Walter, R., Bailly, G., Nischt, M., Alt, F.: Looking glass: a field study on noticing interactivity of a shop window. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '12, pp. 297–306. ACM, New York (2012). http://doi.acm.org/10.1145/2207676.2207718
30. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
31. Robins, B., Dautenhahn, K.: Tactile interactions with a humanoid robot: novel play scenario implementations with children with autism. Int. J. Soc. Robot. **6**(3), 397–415 (2014)
32. Saatci, Y., Town, C.: Cascaded classification of gender and facial expression using active appearance models. In: 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006, pp. 393–398, Apr 2006
33. Sakarkaya, M., Yanbol, F., Kurt, Z.: Comparison of several classification algorithms for gender recognition from face images. In: 2012 IEEE 16th International Conference on Intelligent Engineering Systems (INES), pp. 97–101, June 2012
34. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recogn. Lett. **33**(4), 431–437 (2012). http://www.sciencedirect.com/science/article/pii/S0167865511001607, Intelligent Multimedia Interactivity
35. Smarr, C.A., Mitzner, T., Beer, J., Prakash, A., Chen, T., Kemp, C., Rogers, W.: Domestic robots for older adults: attitudes, preferences, and potential. Int. J. Soc. Robot. **6**(2), 229–247 (2014)
36. Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L.: Gender classification based on boosting local binary pattern. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 194–201. Springer, Heidelberg (2006)
37. Sun, Z., Bebis, G., Yuan, X., Louis, S.J.: Genetic feature subset selection for gender classification: a comparison study. In: IEEE Workshop on Applications of Computer Vision, pp. 165–170 (2002)
38. Tapia, J., Perez, C.: Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of LBP, intensity, and shape. IEEE Trans. Inf. Forens. Secur. **8**(3), 488–499 (2013)
39. Toderici, G., O'Malley, S., Passalis, G., Theoharis, T., Kakadiaris, I.: Ethnicity- and gender-based subject retrieval using 3-d face-recognition techniques. Int. J. Comput. Vis. **89**(2–3), 382–391 (2010)

40. Ullah, I., Hussain, M., Muhammad, G., Aboalsamh, H., Bebis, G., Mirza, A.: Gender recognition from face images with local WLD descriptor. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 417–420, Apr 2012
41. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
42. Ylioinas, J., Hadid, A., Pietikäinen, M.: Age classification in unconstrained conditions using LBP variants. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1257–1260, Nov 2012
43. Ylioinas, J., Hadid, A., Hong, X., Pietikäinen, M.: Age estimation using local binary pattern kernel density estimate. In: Petrosino, A. (ed.) ICIAP 2013, Part I. LNCS, vol. 8156, pp. 141–150. Springer, Heidelberg (2013)

# Comparison of Facial Alignment Techniques: With Test Results on Gender Classification Task

Tunç Güven Kaya(✉) and Engin Firat

Adoniss Software Ltd., Middle East Technical University Technopolis, Ankara, Turkey
{tunc.kaya, engin.firat}@adonissyazilim.com

**Abstract.** In this paper, different facial alignment techniques are revised in terms of their effects on machine learning algorithms. This paper, investigates techniques that are widely accepted in literature and measures their effect on gender classification task. There is no special reason on selecting gender classification task, any other task could have been chosen. In audience measurement systems, many important demographics, i.e. gender, age, facial expression, can be measured by using machine learning algorithms. Moreover; in such a system, these demographics are so substantial since their discriminative features on customers' habits. Due to this importance, any performance enhancement on any machine learning algorithm becomes important. A carefully chosen alignment method can boost performance of machine learning algorithms used within system.

## 1 Introduction

This paper aims to provide a comparison among popular alignment algorithms in terms of their aligning qualities. Since most of the video analytics approaches like age classification, face recognition etc. use facial alignment techniques, it is significant to have a successful and robust alignment technique. As the name 'aligning quality' implies, a quantifying process is required to have an objective criterion. That is why, we test different alignment techniques on gender classification task to measure their qualities relatively.

Gender classification, age group classification, face recognition and expression analysis are mostly used in video analytics for audience measurement systems. All these techniques use some machine learning algorithms in common such as support vector machines, neural networks, decision trees etc. And, for a machine learning algorithm to perform better, the data should be 'good'. At this point, the reader may ask what 'the good data' is. In image processing domain, the good data stands for the data which is decontaminated from the effects of illumination, rotation changes, camera noise etc. There are many approaches to eliminate these side effects listed above. In a similar vein, Gaussian filter, median filter and lots of other spatial filters are widely used to eliminate the camera noise. Likewise, histogram equalization, Tan-Triggs method [1] etc. can be used to normalize variable illumination conditions. Moreover, facial alignment techniques are used in order to normalize face in spatial domain. It is


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811 pp. 86–96, 2014.
DOI: 10.1007/978-3-319-12811-5_6

important for a machine learning algorithm to fit every discriminative feature, i.e. eyes and nose location, to nearly same spatial position. In this paper, we focus on the normalization of head pose; in other words, we will investigate the most common techniques that are used for the removal of the effects of head pose changes.

Before starting such an attempt we first need to understand how pose of head changes. A face might occur in countless poses and instances on a video stream. However, we need to limit our research to the output of the face detectors. This is because a typical pipeline for a class prediction task starts with a 'Face Detection' phase, as it can be seen in (1). Most face detectors lack the ability of capturing faces in various poses such as profile faces, they can detect upright frontal and near-frontal faces. That is the reason why we only focus on these face occurrences in this study.

$$\text{Face Detection} \rightarrow \text{Alignment} \rightarrow \text{Representation} \rightarrow \text{Classification} \quad (1)$$

The face detection part of the pipeline is more or less the same through different applications, since most of the researchers and companies use cascaded classifier [2] or its variations [3]. However, there are lots of debates still going on both in the classification and representation (feature extraction) phases. Support vector machines [4], Random Forests [5], Boosting [6], Bayesian networks [11] and lots of other machine learning algorithms were proposed for the classification task and still there is no agreement. In representation phase, it is also the case. There is a huge body of researches regarding the use of different feature sets [12–14].

What we propose to do in this paper, is to investigate the effects of different alignment algorithms by using same settings for the rest of the pipeline, i.e. detection, representation, classification; and then measure their success rates relatively on a gender classification task. Further details about the gender classification task and the experiment will be given in Sect. 4.

Before going any further, some concepts and terms should be clarified first. The term alignment is used for the process of normalizing the images of the object of interest by eliminating the differences related to in-plane and out-of-plane rotations. And, cropping consists of setting certain boundaries – with a certain heuristic- to the aligned object and taking the area within these boundaries. We will mainly be talking about alignment techniques, but we also provide the cropping heuristics accompanying each alignment technique. Facial landmarks (or facial feature points) refer to somehow important points on face such as eye corners, nostril etc. the term is widely used in the literature.

We aim to explain four different algorithms which are more or less the prototypes of different alignment approaches. Three of them are based on facial landmark (or feature point) extraction and the last one uses statistical methods. The rest of the paper is organized as following: Sect. 2 describes the feature based algorithms while Sect. 3 is about machine learning approach. In addition, Sect. 4 gives the experimental results of different alignment techniques on gender classification task. Finally, the last section is about conclusion and future works.

Note that, this study does not aim to be an exhaustive investigation of all the available alignment techniques, we only examine the most popular algorithms and the results will be given accordingly. Moreover, we aim to create a solid standpoint for future studies.

## 2 Alignment According to Feature Points

First thing that comes to mind is detecting important points on the face and performing alignment accordingly. There exists a huge body of researches on this topic, and the works are so various that should be investigated under a couple of classes.

In this paper, three different alignment algorithms that use facial feature points will be examined, since there are no generally accepted names for the most of these algorithms it was considered to be convenient to give names for the ease of comprehension in the following of this paper.

### 2.1 Two Points Based Alignment

In theory, if one knows the positions of two points on the face image and their reference relations on an upright face, then it is possible to align the face according to the angle between the reference line and the line connecting those two points on the given face image. This approach is used mostly after the detection of the eye centers, since it is relatively easier to detect the eye centers and the relation between the eyes - which is assumed to be parallel to the horizontal axis- can be regarded as a generic relation through different persons.

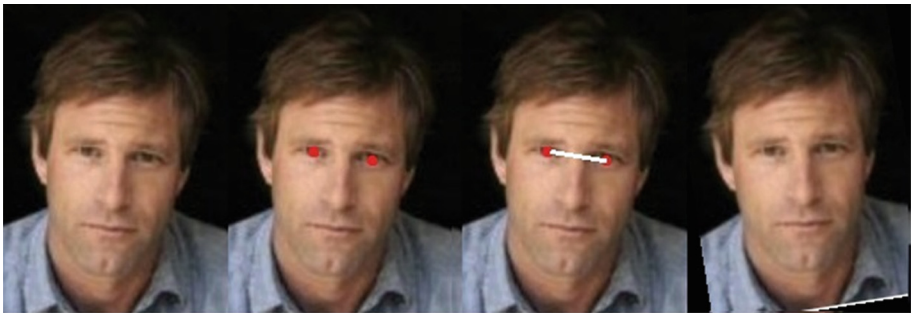


**Fig. 1.** The first column shows the original image taken from LFW database, the second column is the ASEF output, and the third column is the manipulation of the ASEF output. The last column is the aligned face.

In our research, eye points were also selected for the alignment. As mentioned above, any other two points on the face can be used for this technique and the eyes were chosen for their ease of detecting and generalization capabilities. The eyes were detected by using a method called Average of Synthetic Exact Filters (ASEF) [15]. The details of the ASEF algorithm is skipped here, but the researchers who are interested can find the details both in the original paper, the references part and on the internet documentations of its applications which are many.

The output of this method are two points on the image corresponding to the left eye iris, and the right eye iris respectively. It is assumed that the eyes are at the same height on a normalized face, so the alignment process can be performed to straighten the line

between the iris points on the horizontal axis. Figure 1 shows a sample output of the ASEF method and the alignment procedure on an image.

There are two main problems of the two points based alignment algorithms. First, the algorithm can only handle in-plane rotations. Second, even if the two points are detected perfectly and the face is aligned accordingly, the cropping side is still problematic since there is only one distance information available. Using only the distance between the eyes (or any other two points) is not enough to correctly crop the face. For example, some people have higher eye distances than others and vice versa. This causes the cropping algorithm to fail, since, normalization and cropping due to one distance is impossible. Figure 2 shows some examples of the outputs of the ASEF alignment, in which the eyes were detected correctly, but the cropping heuristics failed due to the personal differences of face proportions and head rotations. As you can see the eyes are at the same location, but the noses, mouths and other face parts are at different locations for each sample.

**Fig. 2.** Eyes are correctly found in each sample as you can understand from their fixed locations. However, personal differences and head pose changes led to poorly cropped faces.

### 2.2 Three or More Points Based Alignment

Two points based alignment algorithms can only deal with in-plane rotation. However, in audience measurement applications faces often appear with both out-of-plane and in-plane rotations at the same time. This case is also true for output of face detection algorithms. Hence an alignment system which can handle these rotations simultaneously is needed.

At least three points are required to manage both with in-plane and out-of-plane rotations. In theory, if the reference relation is known among three or more points on a two-dimensional plane, the plane can be rotated in x, y and z axes.

In face related applications, the feature points around eyes, nose and mouth are mostly used [15–17]. These points are detected with a feature detector and brought to fixed places on a model by using an appropriate transformation. Before applying the required transformation, a reference model in which every detected point has a fixed location should be created. The reference model can be formed both empirically and

statistically. Afterwards, the transformation matrix can be calculated to bring the related points to their fixed locations.

In this paper, outputs [25] of a commercial alignment algorithm [8] is used. This algorithm detects seven feature points including four eye corners, two mouth corners and the tip of the nose. Then, applies similarity transformations to bring these points to their fixed locations on a reference model. The basics of the algorithm is shown in Fig. 3, a better and more detailed explanation can be found in [8].

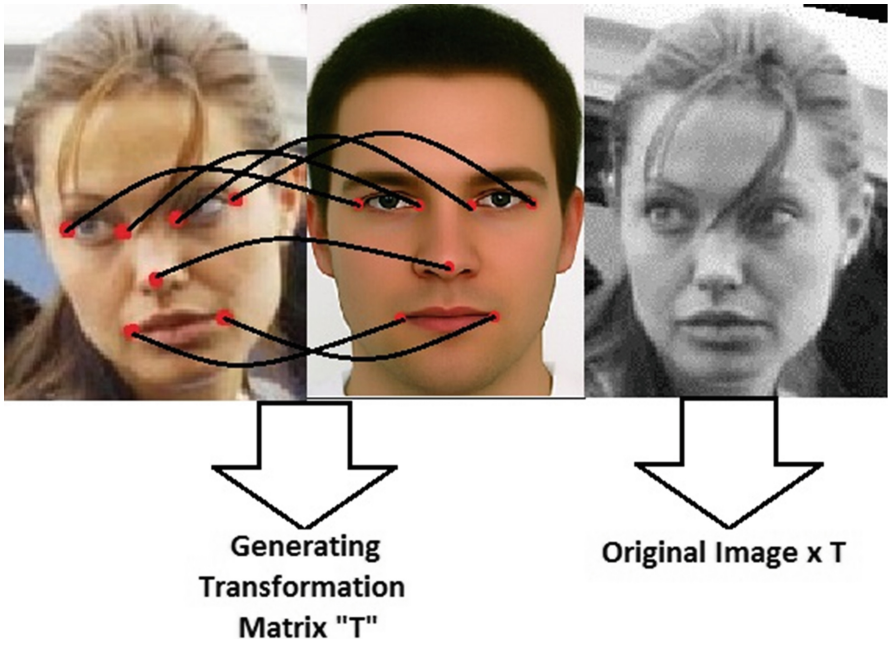



**Fig. 3.** The first image is the original image, the second is the reference model and the third one is the aligned version of the first image. At first, the eye corners, the mouth corners and the nose tip are detected. The red points demonstrate the detected points and the black lines (curvatures) illustrate the correspondences between the original image points and reference model points. Afterwards, the transformation matrix is calculated by comparing these points with their corresponding ones in the reference model. By multiplying the original image with the transformation matrix, we generate aligned face image. The first and the third images are taken from the LFW and the LFW-a databases respectively.

### 2.3 Active Shape Model Alignment

Active shape models for face analysis was proposed by Tim Cootes and Chris Taylor in 1995 [21] and have been quite popular in facial feature detection and normalization domains. There are lots of improvements to the original paper both from Tim Cootes et al. and other researcher groups. Active Appearance Model (AAM) [22] and active appearance models for profile faces [23] are a couple of these improvements. However we are not going to go any further on these improvements here.

The idea behind shape models is creating a statistical face model with labeled data, in which the feature points are annotated manually on large number of face images, and using that model to detect feature points on a new face image. STASM [31] which is an open source application of active shape models was used in this paper and the following figures and results are given accordingly.

Delaunay triangulation [29] method performs a triangulation between points on a two dimensional plane in a way that no single point is in another triangle's circumcircle.

There are couple of applications [32, 33] that use Delaunay triangulation to connect facial feature points and affine warping to normalize the triangles.

Before using this algorithm, a reference model should be created to generate initial Delaunay triangles. The average male face [30] was used to generate the reference model in this study. The reference model and the details of the algorithm can be seen in Fig. 4.

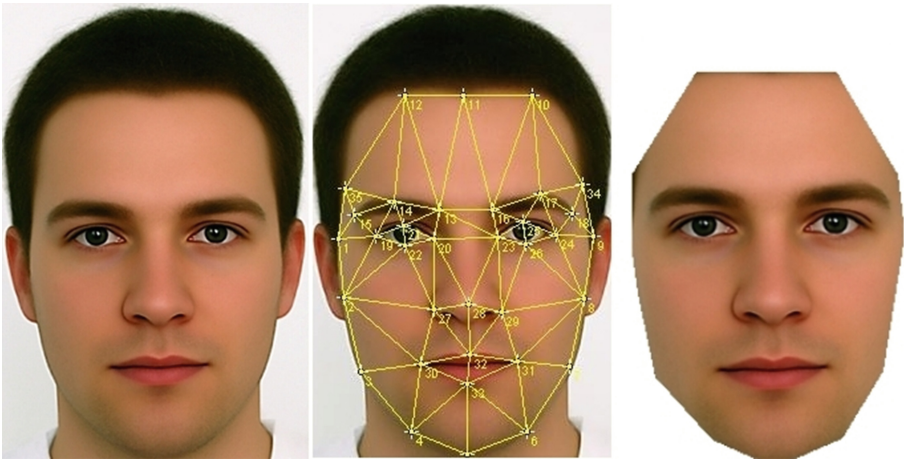


**Fig. 4.** The first column is the average male model. In the second column, the output points of STASM and corresponding Delaunay triangles can be seen. The last column is the cropped version. This model is also used as our reference model. Note that the STASM output and the Delaunay triangulation stage is simplified for the sake of comprehension, normally there are 77 points.

After the creation of the model, for each face image, the following stages are executed in a consecutive manner;

1. Find the facial points by using active shape model or its variants.
2. Create each triangle that corresponds to the ones in the reference model created before.
3. Apply warping transformations to each Delaunay rectangle to fit them into the model.
4. Create the bounding rectangle of the points and crop the image according to that bounding rectangle.

## 3 Machine Learning Based Alignment Techniques

There is another branch of image alignment techniques which does not depend on extraction of features. The method is called congealing, it was first introduced by Viola et al. in 2000 [20] for letter recognition task and has been used widely since. The idea is minimizing the pixel wise variances to decrease the entropy of the given set of images. The algorithm requires a set of roughly aligned images (for example face detector's output) of a certain entity (letter, face, car etc.) and applies transformations like rotation, shearing, translation in x and y dimensions to reduce the variance of the pixels.

There are couple of applications of congealing in facial image alignment domain. Here, we introduce only one of them, and the results will be given accordingly. The algorithm we examine here is called Deep Funneling [10] and it is the successor of the Funneling [9] algorithm. Both algorithms were introduced by Huang and Deep Funnel was chosen because of its superiority in face recognition task [27]. The aligned versions of LFW database are available both for Funneling and Deep Funneling. Deep Funneled version of the LFW database is used in our experiments.

## 4 Results on Gender Classification Task

For comparison purposes, an experiment was conducted on publicly available Labeled Faces in the Wild (LFW) database [7]. Since, different versions of this database - aligned with some of abovementioned algorithms- are available, we thought that it would be a convenient environment to conduct the gender classification test.

LFW database consist of 13233 faces belonging to 5749 different people. Almost all faces in the database are frontal or near frontal and there is a huge variance in age and ethnicity which makes it a good environment for a gender classification task, especially for the ones who want to run their systems in the wild.

First, the database was manually divided into two classes according to the gender. Afterwards, 922 samples were selected from each class in a way that there are no multiple occurrences of the same person. As mentioned above, commercial [25] and deep funneled [26] versions of the database were already available. Therefore, the same samples were selected from those versions of the database as well. The two points based aligned and active shape model based aligned versions of the database were also generated.

After the alignment phase, the databases were cropped with similar heuristics mostly from the corner of the forehead to the chin, note that the active shape model alignment does not need further cropping. Table 2 shows the characteristics of the cropping heuristics, and generated image sizes for each alignment algorithm. After the cropping phase, the output images were scaled to a certain size which is different for each alignment technique. Sizes were selected in a way that total pixel count for each technique is more or less the same. This way, nearly the same amount of data is provided to machine learning algorithms. Moreover, for the testing purposes, the output of the OpenCV [29] face detector is included as a reference which corresponds to no alignment condition.

The misaligned or poorly cropped samples were eliminated from the training data, but kept in the test data. The reason behind is that, the researchers have the chance to

manipulate (eliminate the poor ones) the training data, but there is no way of changing the poor test data which is coming from the real time video, so we thought it would be a more realistic experiment environment. After the elimination phase, the training databases of different alignment algorithms were equalized in number (850 samples per class) to avoid any kind of bias due to statistical learning phase.

The results are given at Table 1, and the discussion is on the last section.

**Table 1.** The heuristics for cropping are given for each alignment algorithm. The output sizes are arranged in a way to balance the pixel numbers (and feature numbers for machine learning) of different algorithms.

| Algorithm | Cropping heuristic | Size of the output |
|---|---|---|
| Face detector | No process, output of detection. | 64 × 64 |
| Two points based | see Fig. 5 | 54 × 81 |
| Three or more points based | see Fig. 5 | 55 × 77 |
| Active shape model based | No process, cropping heuristic is built-in | 54 × 81 |
| Deep-funneled | see Fig. 5 | 54 × 81 |

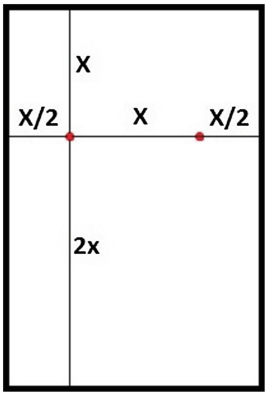



**Fig. 5.** The red dots are the eye centers and 'x' is the eye distance. The width of the cropped area is 2*x and the height is 3*x.

**Feature Selection and Machine Learning.** LBP [19] features have been used in many researches [12, 18] and have good results on gender classification task. That is why we used LBP features for representation phase here, but other methods could also be used.

Two different machine learning algorithms were used and the results are given accordingly. Since, our aim is not the comparison of machine learning tools, we do not investigate their relative success rates and other aspects further.

**Table 2.** The results of the gender classification task. Random forests algorithm was used with 100 trees and 50 features per tree. AdaBoost was used with 100 iterations (weak classifiers). The last column shows the results of the experiment in which the whole database was given as the test.

| Machine learning algorithm | Alignment technique | 10-Fold | Full LFW test |
|---|---|---|---|
| Random forests | Face detector | 74.80 | 67.64 |
| | Two points based | 80.14 | 78.07 |
| | Three or more points based | 88.56 | 83.24 |
| | Active shape model based | **91.63** | **85.72** |
| | Deep funneled | 89.05 | 82.59 |
| AdaBoost | Face detector | 76.24 | 68.92 |
| | Two points based | 79.29 | 77.86 |
| | Three or more points based | 89.10 | 82.51 |
| | Active shape model based | **88.89** | **84.36** |
| | Deep funneled | 89.48 | 80.94 |

## 5 Conclusion

All of the alignment techniques performed better than no alignment condition on the gender classification task. The best success rate was achieved by using active shape models and Delaunay triangulation method. However, three or more points based and machine learning based techniques have quite similar test results. Two points based methods have relatively poor results, which might be due to the problems mentioned in the related section. The results are consistent with the finding of [24] which was conducted on the same database.

Experiments done for gender classification task, but as already mentioned these results are applicable to other tasks which needs machine learning algorithms to run. For example, tasks like age detection, facial expression recognition, and gender recognition all use machine learning algorithms. Prediction over classes (i.e. gender classes: male or female) are done over a subset of extracted features which is discriminative for that class prediction. These discriminative features belong to some specific facial parts. Hence, alignment of face, in other words normalization, becomes a crucial part to enhance performance of prediction related tasks. This is why gender classification task is affected by alignment methods.

In this work, widely accepted improvement methods for alignment phase were explained and their effect on gender classification task was measured. Active Shape Model based alignment technique outperformed all other alignment techniques. Due to the nature of the general pipeline for prediction alike tasks, one can estimate that success rates of other tasks like age prediction, facial expression recognition, and facial recognition can be increased by using an Active Shape Model based alignment technique.

## References


1. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process. **19**(6), 1635–1650 (2010)
2. Viola, P.A., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of IEEE Computer Society Conference on Computer Vision Pattern Recognition, vol. 1, pp. 511–518 (2001)
3. Zhang, L., Chu, R., Xiang, S., Liao, S., Li, S.Z.: Face detection based on multi-block LBP representation. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 11–18. Springer, Heidelberg (2007)
4. Cortes, C., Vladimir, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Freund, Y.: Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
7. Huang, G.B., et al.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)
8. Wolf, L., Hassner, T., Taigman, Y.: Effective face recognition by combining multiple descriptors and learned background statistics. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **33**(10) (2011)
9. Huang, G.B., Mattar, M.A., Lee, H., Learned-Miller, E.G.: Learning to align from scratch. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS (2012)
10. Huang, G., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: IEEE International Conference on Computer Vision (2007)
11. Jensen, F.V.: An Introduction to Bayesian Networks. Springer, New York (1996)
12. Lian, H.-C., Lu, B.-L.: Multi-view gender classification using local binary patterns and support vector machines. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 202–209. Springer, Heidelberg (2006)
13. Jabid, T., Kabir, H., Chae, O.: Gender classification using local directional pattern (LDP). In: 2010 International Conference on Pattern Recognition (ICPR 2010), pp. 2162–2164 (2010)
14. Wang, J.G., Li, J., Yau, W.Y., Sung, E.: Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. In: Proceedings of CVPR, San Francisco, pp. 96–102 (2010)
15. Bolme, D.S., Draper, B.A., Beveridge, J.R.: Average of synthetic exact filters. In: CVPR (2009)
16. Wang, P., Green, M., Ji, Q., Wayman, J.: Automatic eye detection and its validation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 164–171 (2005)
17. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: CVPR (2012)
18. Huang, D., Shan, C., Ardabilian, M., Wang, Y., Chen, L.: Local binary patterns and its application to facial image analysis: a survey. IEEE Trans. Syst. Man Cybern. Part C **41**(6), 765–781 (2011)
19. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recogn. **29**(1), 51–59 (1996)
20. Miller, E., Matsakis, N., Viola, P.: Learning from one example through shared densities on transforms. CVPR **1**, 464–471 (2000)

21. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models—their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
22. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE TPAMI **23**, 681–685 (2001)
23. Cootes, T.F., Wheeler, G.V., Walker, K.N., Taylor, C.J.: View-based active appearance models. Image Vis. Comput. **20**(9–10), 657–664 (2002)
24. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recogn. Lett. **33**(4), 431–437 (2012)
25. Labeled Faces in the Wild-a (LFW-a). http://www.openu.ac.il/home/hassner/data/lfwa/
26. Deep Funneled LFW. http://vis-www.cs.umass.edu/lfw/#deepfunnel-anchor
27. Face Recognition Task on LFW. http://vis-www.cs.umass.edu/lfw/results.html
28. Bradski, G.: The OpenCV Library. Dr. Dobb's J. Softw. Tools (2000)
29. Delaunay Triangulation, http://en.wikipedia.org/wiki/Delaunay_triangulation
30. Average male and female faces. http://www.uni-regensburg.de/
31. Milborrow, S., Nicolls, F.: Active shape models with SIFT descriptors and MARS. VISAPP **1**(2), 5 (2014)
32. Tanigman, Y., Yang. M, Ranzato, M., Wolf, L.: DeepFace: Closing the gap to human-level performance in face verification (2014)
33. Irfanoglu, M., Gokberk, B., Akarun, L.: 3d shape-based face recognition using automatically registered facial surfaces. In: Conference on Pattern Recognition, vol. 4, pp. 183–186 (2004)

# Multi-view Face Detection with One Classifier for Video Analytics Systems

Tunç Güven Kaya(✉) and Engin Firat

Adoniss Software Ltd., Middle East Technical University Technopolis, Ankara, Turkey
{tunc.kaya,engin.firat}@adonissyazilim.com

**Abstract.** In a video analytics for audience measurement system, dwell time, gaze, and opportunity-to-see statistics are required most of the time. To generate these statistics, more than one face detector is used in order to capture both profile and frontal faces. In this paper, we present a novel approach for face detection in video analytics. The assumption is that; the face occurrences are limited in such systems and one classifier is able to capture all of these occurrences. By using MB-LBP for feature extraction and Gentle Boost for statistical learning we trained a classifier which is able to detect both profiles and frontal faces with more than 80 % success rate. The proposed system also uses an efficient scanning algorithm and achieves 60 fps on a 720p video.

**Keywords:** Multi-view face detection · Object detection · Audience measurement system · Multi-block lbp · Video analytics

## 1 Introduction

Face detection, tracking, pose estimation and machine learning associated techniques (such as gender detection, age classification) can be considered as the building blocks of a video analytics for audience measurement system. By using these building blocks, demographic data, hot-zone, dwell time, attention time, and opportunity-to-see statistics can be generated.

In a traditional video analytics system, three face detectors are used for updating dwell time and opportunity-to-see statistics, because, the system should capture both left and right profiled faces. Since, running the face detector is computationally expensive the use of these methods might merge bottlenecks in the system.

What we propose in this paper is a profile and frontal face detection system by using just one classifier. This new system is able to run at 60 fps on a 720p video when combined with the efficient scanning algorithm which is also explained later in this paper. Both the detection rates and the execution time of the system are competitive as can be seen in the results section. Moreover, the system is able to find nearly all possible face occurrences in a video analytics for audience measurement system.

The paper will continue as following: Sect. 2 investigates former solutions to a similar problem, Sect. 3 explains proposed method in a detailed manner, Sect. 4 gives experimental results and, Sect. 5 is about conclusion and advantages of the proposed system.


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811 pp. 97–108, 2014.
DOI: 10.1007/978-3-319-12811-5_7

## 2 Related Works

Face detection task has been the topic of a lot of researches. However, at some time the research of Viola and Jones [10] became a milestone for single view (frontal) face detection task; and since then the studies have been continuing on multi-view face detection mostly.

The studies mostly focus on detecting every possible face occurrences with multiple classifiers. Here, we will investigate couple of them and explain why these approaches are not appropriate for a video analytics system.

Viola and Jones tried to apply the approach in [10] to multi-view face detection with the new set of Haar-like features [9]. Their idea was to create subsets each covering a certain group of face occurrences and generate detectors for each subset. This approach consisted of two processes, the first was the determination of the pose with a decision tree and the second was the execution of the appropriate detector according to the decision tree's output. In their system, there were 12 different classifiers for in-plane rotation and two for out-of-plane rotation along with their corresponding decision trees for pose-estimation.

Huang et al. [8] proposed to train multiple classifiers for each pose of face and to execute these classifiers on the frame in a coarse-to-fine manner, meaning that at the first level only the first five stages of the cascade classifiers would run and return confidence values, according to the maximum confidence value, the corresponding classifier would execute fully. The main pitfall of this system is its performance. To illustrate a bit further; for all different face poses, a classifier is needed and this classifier's first five stages should be run on the frame. Thus, running all these classifiers causes a performance bottleneck on the measurement.

In both approaches, it is assumed that a face can be found in any angle both in-plane and out-of-plane rotation; and the differences are mostly due to the pose estimation phase which determines the classifier to execute. However, in an audience measurement system, the occurrences of faces are limited to a certain domain. For example, there cannot be +180° in-plane rotated face in a realistic[1] scenario. Therefore, we examined audience measurement videos and decided probable face occurrences manually and Fig. 1 demonstrates our convention of realistic and unrealistic. Hence, our research is narrowed down to those occurrences. In a realistic measurement scenario, facial poses that should be considered contains following in-plane and out-of-plane poses:

In-plane: ± 15° (rotation in roll axis)
Out-of-plane: ± 15° (rotation in pitch axis)
Out-of-plane: ± 90° (rotation in yaw axis).

[1] The term 'realistic' is defined by the restrictions of a specific system in which the camera is set at about 2 m high and directly looks at the customers. Different audience measurement setups might need to cover different face occurrences and the term 'realistic' should be defined from the scratch for those systems.

**Fig. 1.** The first row shows realistic face occurrences while the second row shows unrealistic face occurrences in an audience measurement video.

## 3 Proposed Method

As mentioned above, researchers assume that it is not possible to detect every occurrence of a face with just one classifier. However, in an audience measurement system there are limitations on face occurrences, and actually we do not need to detect every face occurrences. A face, for example, with +180° in-plane rotation is considered unrealistic for such systems.

Moreover, we thought it would be a good research topic to try and see if one classifier is sufficient for the detection of all possible faces in a video analytics system for audience measurement. Our approach to the problem and the results we get will be discussed in the rest of this paper.

In this section, our solution to the multi-view face detection problem will be discussed. There are couple of stages for constructing a multi-view face detector and these stages will be explained clearly for the reader's full comprehension on the subject. The subtopics can be listed as; data collecting, feature generation (the representation of the data), training, and pose-estimation.

### 3.1 Data Collecting

In statistical learning problems, the quantity and the quality of the data are assumed to be the most important aspects. Data selection for the training phase highly affects the quality of the classifier. Moreover, multi-view face detection with a single classifier is already a complicated task by its nature, the faces of different angles must be arranged in a compatible manner. That is why the data collecting (mostly positive data) phase was carefully performed and selected as a section on our research.

To be able to fulfil the training purposes, images were collected from Head Pose Image Database [4] and the internet. The training set consist of 3600 images equally distributed among left profile, right profile and frontal faces. An application was developed to assist in cropping process. With the help of this application, for all different poses, facial areas were cropped in a semi-automated manner and the multi-class (i.e. faces in different orientations) harmony between different pose groups were

taken into consideration. By multi-class harmony, we mean that profile, half-profile and frontal faces must be cropped in a compatible manner; the starting (top left) and end (bottom right) points of the rectangles should more or less match and each rectangle must have the same aspect ratio. Otherwise, the task of the classifier would be much more complicated than it already is. The starting of the hair line was chosen as the top of the rectangle and the tip of the chin was chosen as the bottom of the rectangle. The distance between the hair line and the tip of the chin was also used as the width of the rectangle, so the aspect ratio was set to 1:1. Other heuristics such as changing the aspect ratio, including the hair or the ears etc. and their effects on the classification task might be the topic for a further research. Figure 2 shows some correct and wrong applications of our cropping heuristic.

**Fig. 2.** The faces are cropped according to the red rectangles. The first row shows correct cropping; aspect ratios are the same for each rectangle and the rectangles are well-placed. The second row shows poorly cropped samples; the aspect ratios differ among the rectangles and some of the rectangles are misplaced.

Additionally, to further increase the ability of generalization of the collected data, some post processes were applied on annotated images. At first, a random linear translation was applied on selected rectangle regions within the images. This process shifts selected rectangle areas by random distances in both x and y direction. The upper limit of the rectangle shifts were empirically chosen as 5 pixels in both directions. Different limit values might also be used. With the help of this post process method, trained classifier would not learn restricted face images. In other words, false negatives will not be generated because of the classifier's restrictive sensitivity to small pixel changes. Secondly, a random affine transformation (within the range of [−20, 20] degrees) around x, y, and z axes was performed. By doing so, similar to first preprocess technique, excessive sensitiveness of trained classifier to rotations was prevented. For a final enhancement, all the cropped face images were flipped horizontally and added to the database to avoid any kind of unpredictable bias due to illumination or pose.

By applying these post process methods, collected data gained greater ability to mimic multi-view faces in a real retailer environment scenario. In this way, higher combinations of multi-view facial poses can be captured by the trained classifier.

### 3.2 Feature Extraction

Feature generation is the stage that the data is represented in another domain to ease the task of the machine learning algorithms and improve their success rates. In face detection researches, HOG (histogram of oriented gradients) [11], Haar-like features [2] and LBP (local binary patterns) [5] are the most used techniques. Different algorithms have their pros and cons with regard to speed, resistance to illumination changes, ease of implementation etc. After empirical studies, we decided to use a variant of LBP features. The ease of implementation, extraction speed (time complexity) and the generalization power of LBP features were the key points behind this decision.

**Local Binary Patterns (LBP).** Local binary patterns were first introduced in 1994 and have been used for face recognition, texture classification and object representation areas widely [13]. The idea is coding the texture information into binary (or decimal) string by comparing an anchor pixel with its neighbors. At the end of feature extraction process, each pixel holds information about intensity changes relative to its neighbors. LBP features are discriminative in the manner of relative intensity changes and this discriminative feature can be used for different tasks.

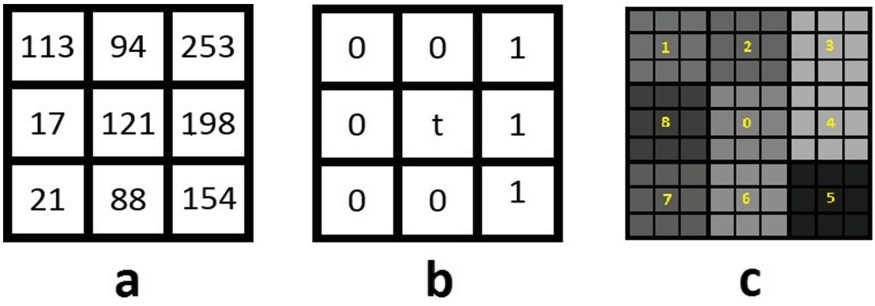



**Fig. 3.** (a) Is a sample neighborhood on an image and the numbers are the gray level intensities of the pixels. (b) Shows (a)'s output as LBP, note that the value of the center pixel is used as a threshold and if the pixel's value is greater than the threshold, the corresponding pixel on LBP image is 1, otherwise 0. The binary value is 00111000 and the decimal value is 58 in this case. (c) Is a sample application of MB-LBP, different sub regions are represented with numbers and with different tones of gray. The algorithm calculates the average gray level intensity for each sub region and performs the same process as (a) and (b).

By noting LBP features' localness, one could understand that they lack the ability of capturing global characteristics of face or related objects. That is why a lot of improvements have been suggested in the image processing domain [7, 14] and Multi-scale Block LBP [7] is one of them. The idea behind MB-LBP is to decrease localness of the features and increase the ability of capturing global characteristics of the object of interest.

**Multi-scale Block Local Binary Patterns (MB-LBP).** The original LBP algorithm is based on pixel-wise comparisons of pixels in a certain neighborhood as can be seen

in Fig. 3. MB-LBP simply takes this idea and applies it to predefined sub regions. The average grey-level intensities of each sub-region are calculated and compared with the one of anchor sub region. In this way, we are capable of generating global features as well as the local features with the use of different-sized sub regions.

### 3.3 Training

In the training phase, the cascaded classifier approach which was introduced by Viola and Jones in 1998 [10] was used. The main idea behind this approach is elimination of background regions within successive classifier layers which are called stages. The complexity of the stages increases gradually, meaning that the number of weak classifiers increase stage by stage so does the execution time. Therefore, rejecting the background regions in earlier levels increases the speed of the classifier. We used AdaBoost [15] algorithm to form the stages, the rest of this section explains the training process with more details.

**Stage Training with AdaBoost.** AdaBoost is a machine learning algorithm proposed by Freund and Schapire in 1995 [15]. The algorithm takes a number of weak classifiers and combine them in a way to form a final strong classifier by arranging their weights and orders. There are different versions - mostly due to the error functions - of AdaBoost algorithms and we empirically chose Gentle Boost for our task. There are two advantages of Gentle Boost; it creates less number of weak classifier which is very essential for our task since it reduces the execution time, and it is more tolerant to the noisy data than the original version. Our primary concern in the training phase was the fast elimination of the background regions within the first stages to avoid high computational costs in the latter stages which have higher numbers of weak classifiers. The precision of the classifier was secondary. Because, the classifier was designed to be a part of an audience measurement system in which faces appear in the frame of interest more than one time unlike still images. The researchers who aim to work on still images should create much more sensitive classifiers to eliminate the case of false negatives.

Before the training phase the face images were resized to 24 × 24 and every possible sub region was created for feature extraction. Afterwards, the corresponding MB-LBP values for each negative and positive sample were created. The output of this process is our labelled data ready to give to the machine learning algorithm. To create a cascade classifier, we need to form every stage one by one by using the preceding stage's false positives as negative data and randomly selecting positive data from the initial positive samples. Figure 4 explains the process with the details and gives a pseudo code.

**Training of the Last Stage - Pose Estimator.** The multi-view face detector is designed to find various posed faces, and the output of this classifier should be interpreted to examine the orientation of the face. This information about the face-pose can be used as a feedback mechanism for the face detector, attention time, dwell time, and opportunity-to-see statistics. Hence, the last stage of the classifier was designed to work as a face-pose estimator.

Face-pose estimation is a multi-class classification problem. The classifier should be able to match the input with one of face orientation classes. For a measurement system that is taking into consideration, three classes which are named as right profile, left profile, and frontal, were sufficient, since these classes are enough for accurate updates on attention time, dwell time and opportunity-to-see demographics. Frontal faces were selected among the faces which have yaw rotations between the range −20° to +20°, left profile faces were selected so that they have yaw rotations +20° to +90° and right profile faces were selected so that they have yaw rotations −20° to −90°. In each class, we aimed to obtain a homogeneous distribution for pitch rotation, i.e. samples have different pitch rotations in range −15° to +15° homogeneously. The rotations in roll axis were not our concern in this labelling.

```
1.  Algorithm for the stage training.
2.  FUNCTION LearnData:
3.          Select positive and negative data subset from supplied samples
4.          Create stage classifier according to the specified ML algorithm, i.e. AdaBoost
5.
6.  FUNCTION Predict:
7.          Run previous stages to predict given sample is a face or not
8.          Label the data according to the original labels (false positives = negative)
9.
10. FUNCTION Merge:
11.         Merge all stages into a single cascade classifier.
12.
13. FOR # of total stage counts:
14.         IF this is first stage:
15.                 DO LearnData
16.         OTHERWISE:
17.                 FOR # of total trained classifier so far:
18.                 Collect positives and false positives by performing DO Predict
19.                 DO LearnData by feeding collected positives and false positives
20. DO Merge
```

**Fig. 4.** Pseudo code for stage training.

For the training phase of the pose estimator, true positive outputs of the trained multi-view face classifier were used. By using test videos which were obtained from real retailer environments, different posed faces were collected and labeled manually as ‘right profile’, ‘left profile’ and ‘frontal’. More than 1500 cropped images were used for the training and the data were equally balanced among these three classes.

For the statistical learning phase, the cropped images were normalized in scale (24 × 24) and LBP features (484 in total) were generated for the representation. The random forests algorithm [6] was used for the learning phase due to its resistance against over fitting and bias. The parameters for the random forests were selected after empirical tests and it was comprehended that increasing the number of trees would not

improve the results after a certain number, in this case 100 trees. Also the depth of the trees should be set somewhere around 5–20 to avoid underfitting and overfitting. The number of features per tree was chosen according to the ratio between the number of features and number of trees. In average, each feature was used in four different trees in this setup. The selected parameters are as following:

Number of trees: 100
Number of features per tree: 20
Maximum depth of a tree: 5.

For the testing, videos - different from the training samples - were collected, and then the multi-view face classifier's results on these videos were given to the pose-estimator as the input. The success rates for "left profile, right profile and frontal" are 0.89, 0.88 and 0.94 respectively.

### 3.4 Efficient Scanning Algorithm

So far, the multi-view face classifier and pose estimator were trained. This work is enough for a well running detector, but there is still some room for performance enhancements. This section is an optional stage for a well running detector, but for the ones who want additional performance (in terms of computation) increase should implement an efficient scanning technique that is about to be explained.

However, before diving into the technical details of this efficient scanning algorithm, let us explain classical method in a few words. In the classical scanning method [10], used in most computer vision libraries, an image pyramid for a given image is constructed by using a scaling factor. Figure 5 is an illustration of such an image pyramid. For all of the pyramid levels, an iterative scanning by using a constant sized (classifier size) rectangle is executed to find faces. This way, faces can be found within all supplied still images to detector. Normally image pyramid construction and scanning on pyramid levels are computationally intensive tasks, and repeating these tasks from scratch for all supplied still images might emerge a performance bottleneck.

Thanks to its feature extraction method and less number of weak classifiers, our trained multi-view face detector already runs in near real-time (15 fps) on a $1280 \times 720$ (see Table 3 for details) full frame video. However, finding the same face 15 times in a second does not have much effect on the quality of the system. Performance of machine learning algorithms (such as gender detection, age classification) and face trackers is not affected by feeding that much information to these algorithms. This situation gives a possibility to develop techniques which decrease processing density.

There are couples of ways to achieve this. Here, two of them will be stated. The first one called "Honeycomb" [3] narrows the search space down, and the second one [2] reduces the computational processes by processing only a certain portion of the image pyramid.

Honeycomb structure was suggested by Ernst and Küblbeck in [3]. The idea is to reduce the density of search space by leaving out some of the regions at the first iteration. The image is divided into honeycombs of different levels inheriting a coarse-to-fine approach, and at first iteration, only the first level which is the coarsest is classified. If the confidence of that region is above a threshold then the denser search

would take place, around that region, on the second and the third levels in a similar manner. This approach was not used in this paper since it requires the re-training of the classifier in order to get the confidence values. Nevertheless, it is considered as a part of a further research.

The second approach which was introduced by Küblbeck and Ernst in [2] reduces the computational density by processing only a certain part of the image pyramid on each frame. This part is chosen according to the frame number and the details can be seen both on the Fig. 5 and Table 1. The modulus of the frame number is used for the

**Table 1.** The visualization of how the computation load is distributed almost homogenously among different ticks of the system. Here, each color denotes a different system tick. It should be noted that, in each tick, total computations are arranged to be nearly equal. The table is given for a 1280 × 720 video with the scale factor of 1.25, different configurations can be arranged with similar manner. The results under the "total loop count" column are not exact calculations, but approximations. This configuration is more or less the same with the original paper [2].

| Pyramid Level | Width | Height | Total Loop Count | Tick count, in which the computation performed |
|---|---|---|---|---|
| 1st | 1.280 | 720 | 921.600 | 1st tick / 2nd tick |
| 2nd | 1.024 | 576 | 589.824 | 3rd tick |
| 3rd | 819 | 461 | 377.487 | 4th tick |
| 4th | 655 | 369 | 241.592 | |
| 5th | 524 | 295 | 154.619 | 5th tick . . . |
| 6th | 419 | 236 | 98.956 | |
| 7th | 336 | 189 | 63.332 | |
| 8th | 268 | 151 | 40.532 | |
| 9th | 215 | 121 | 25.941 | |
| 10th | 172 | 97 | 16.602 | |
| … | … | … | … | |
| | | | 2.555.048 | |

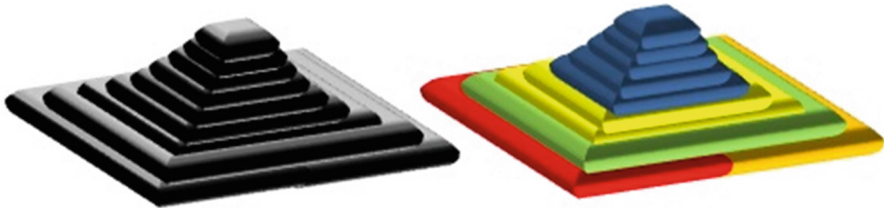


**Fig. 5.** The image pyramid of the original scanning algorithm [10] is shown on the left, note that all of the layers are processed at the same tick. The scanning algorithm used in this paper is shown on the right, each color corresponds to the tick number in which the layer(s) (of that color) is processed. The figure was created in a way that the sizes of the rectangular areas correspond to the real ratios, as you can see each color covers more or less the same area. For further investigation, Table 1 can be examined.

determination of the processing unit. The system continues to take the frames, but only a certain subset of the current frame is processed. In this study, the image pyramid was divided into 5 subsets so that each subset has more or less the same number of processes. Different heuristics might be used for the generation of the subsets for different resolutions and scaling factors.

## 4 Experimental Results and Analysis

The detection rates of the classifier were tested on CMU + MIT database [18] (for frontal faces) and the Sheffield Face Database [17] (for profile faces). The detection rates and false positive counts can be seen in Table 2.

**Table 2.** Success rates and false positive counts of the proposed classifier on CMUMIT database [12] for frontal face detection (only the A, B and C tests of CMUMIT database were used.). The profile test was conducted on Sheffield Face Database.

| | Frontal faces | | Profile faces | |
|---|---|---|---|---|
| | Success rate | False alarm | Success rate | False alarm |
| Proposed Method | %84.63 | 112 | %81.17 | 68 |
| Viola-Jones [10] | %90.8 | 95 | - | - |
| Rowley [17] | %89.2 | 95 | - | - |
| Zhang-Chu [14] | %90.7 | 57 | - | - |

We also performed a test to measure the execution time of the classifier in different environments. For this purpose we have found eleven videos from the internet which encapsulate a couple of real retailer scenarios, such as retailers with complex backgrounds, different lightning conditions and crowded scenes. All classifiers were executed on these videos and the average execution times are given in Table 3. Experimental setup consists of an Intel Core i7 3630QM processor with a 2* 8 GB dual channel RAM. This processor has 8 cores for processing purposes and during the test all the cores are used. The proposed face detector - without the efficient scanning algorithm which was mentioned in subsection 3.4 - processes a 720p frame in less than 60 ms. It is much (three to four times) better than Haarlike features based face detectors of OpenCV and just a little bit slower than OpenCV LBP detectors. The scaling factor which illustrates the scaling ratio of the image pyramid was set to 1.25 in the test phase. Denser or sparser searching configurations might also be applied. However, the proportions among the results would not change due to the fact that the algorithm complexity is independent of the size of the levels in the image pyramid.

As you can see, the success rates of the proposed method is behind the other methods, however it manages to find both frontal and profile faces with more than 80 %

**Table 3.** Comparison of performance values of different classifiers. The first row shows the results of the proposed classifier and the remaining rows show the classifiers available in OpenCV [16], the classifier names are self-explanatory so a further explanation is avoided. The tests were performed on 11 different 720p HD videos to get healthier results, and the scaling factor was set to 1.25.

| Classifier Name | Captures | Execution Time |
|---|---|---|
| Proposed Classifier | Both profiles and frontal faces | 58.193 ms |
| Haarcascade_frontalface_alt2 | Only frontal faces | 160.9 ms |
| Haarcascade_frontalface_alt | Only frontal faces | 170.028 ms |
| Haarcascade_frontalface_default | Only frontal faces | 201.241 ms |
| Lbpcascade_frontalface | Only frontal faces | 43.29 ms |
| Lbpcascade_profile | Only right profile faces | 44.159 ms |
| Haarcascade_profile | Only right profile faces | 192.955 ms |

success by using just one classifier. Moreover, the execution time is very promising. The fastest classifier in the list (frontal LBP classifier) requires two more similar classifiers to capture left and right profiled faces, and the time gets triple.

## 5 Conclusion

In this paper, a new face detection approach for video analytics systems is proposed and some possible performance enhancements are defined. Proposed detector finds faces that can have different orientations such as left-right half-full profile faces and frontal faces. Due to classifier's nature (fast rejection of the background regions and feature type), detector can achieve considerable computational performance value which is appropriate for an audience measurement system. In addition, by implementing a new scanning technique, computational processing density of the detector can be decreased further. The proposed face detector is able to run at 60 fps on a 720p video when combined with the abovementioned scanning algorithm.

Proposed face detector not only have enhancements in terms of computational performance but also have some novelties in terms of its usability. It can be used for updating dwell time, gaze and opportunity-to-see statistics which requires additional effort in the former approaches. The system is also appropriate for digital signage tasks since the computational burden is relatively less.

## References

1. Jones, M., Viola, P.: Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96, vol. 3 (2003)
2. Küblbeck, C., Ernst, A.: Face detection and tracking in video sequences using the modified census transformation. Image Vis. Comput. **24**, 564–572 (2006)

3. Ernst, A., Küblbeck, C.: Fast face detection and species classification of African great apes. In: 2011 IEEE International Conference on Advanced Video and Signal based Surveillance (2011)
4. Gourier, N., Hall, D., Crowley, J.L.: Estimating face orientation from robust detection of salient facial structures. In: FG Net Workshop on Visual Observation of Deictic Gestures. Cambridge, UK: FGnet (IST–2000–26434) (2004)
5. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
6. Breiman, L.: Random Forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Zhang, L., Chu, R., Xiang, S., Liao, S., Li, S.Z.: Face detection based on multi-block LBP representation. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 11–18. Springer, Heidelberg (2007)
8. Wu, B., Ai, H.Z., Huang, C., Lao, S.H.: Fast rotation invariant multi-view face detection based on real adaboost. In: FG (2004)
9. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: ICIP (2002)
10. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Press, Los Alamitos (2001)
11. Dalal, N., Triggs, B.: Histogram of oriented gradient for human detection. In: CVPR (2005)
12. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 23–38 (1998)
13. Ojala, T., Pietikainen, M., Maeenpaa, T.: Multiresolution gray-scale androtation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24** (7), 971–987 (2002)
14. Guo, Z., Zhang, L., Zhang, D., Mou, X.: Hierarchical multiscale LBP for face and palmprint recognition. In: International Conference on Image Processing, Hong Kong, pp. 4521–4524 (2010)
15. Freund, Y.: Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
16. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
17. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**(1), 23–38 (1998)
18. Schneiderman, H., Kanade, T.: A statistical model for 3D object detection applied to faces and cars. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, June 2000

# Modelling Consumer Behaviour

# Online Audience Measurement System Based on Machine Learning Techniques

Vladimir Khryashchev(✉), Andrey Priorov, and Alexander Ganin

P.G. Demidov Yaroslavl State University, Yaroslavl, Russia
dcslab@uniyar.ac.ru

**Abstract.** An application for video data analysis based on computer vision methods is presented. The proposed system consists of five consecutive stages: face detection, face tracking, gender recognition, age classification and statistics analysis. AdaBoost classifier is utilized for face detection. A modification of Lucas and Kanade algorithm is introduced on the stage of tracking. Novel gender and age classifiers based on adaptive features, local binary patterns and support vector machines are proposed. More than 92 % accuracy of viewer's gender recognition is achieved. All the stages are united into a single system of audience analysis. The system allows to extract all the possible information about depicted people from the input video stream, to aggregate and analyze this information in order to measure different statistical parameters.

**Keywords:** Online audience measurement · Machine learning · Gender classification · Age estimation · Local binary patterns · Support vector machines

## 1 Introduction

Automatic video data analysis is a very challenging problem. In order to find a particular object in a video stream and automatically decide if it belongs to a particular class one should utilize a number of different machine learning techniques and algorithms, solving object detection, tracking and recognition tasks [1–6]. A lot of different algorithms, using such popular techniques as principal component analysis, histogram analysis, artificial neural networks, Bayesian classification, adaptive boosting learning, different statistical methods, and many others, have been proposed in the field of computer vision and object recognition over recent years. Some of these techniques are invariant to the type of analyzed object, others, on the contrary, are utilizing aprioristic knowledge about a particular object type such as its shape, typical color distribution, relative positioning of parts, etc. [7]. In spite of the fact that in the real world there is a huge number of various objects, a considerable interest is being shown in the development of algorithms of analysis of a particular object type – human faces. The promising practical applications of face recognition algorithms can be automatic number of visitors calculation systems, throughput control on the entrance of office buildings, airports and subway; automatic systems of accident prevention, intelligent human-computer interfaces, etc.

Gender recognition, for example, can be used to collect and estimate demographic indicators [8–10]. Besides, it can be an important preprocessing step when solving the


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811 pp. 111–122, 2014.
DOI: 10.1007/978-3-319-12811-5_8

problem of person identification, as gender recognition allows twice to reduce the number of candidates for analysis (in case of identical number of men and women in a database), and thus twice to accelerate the identification process.

Human age estimation is another problem in the field of computer vision which is connected with face area analysis [11]. Among its possible applications one should note electronic customer relationship management (such systems assume the usage of interactive electronic tools for automatic collection of age information of potential consumers in order to provide individual advertising and services to clients of various age groups), security control and surveillance monitoring (for example, an age estimation system can warn or stop underage drinkers from entering bars or wine shops, prevent minors from purchasing tobacco products from vending machines, etc.), biometrics (when age estimation is used as a part that provides ancillary information of the users' identity information, and thus decreases the whole system identification error rate). Besides, age estimation can be applied in the field of entertainment, for example, to sort images into several age groups, or to build an age-specific human-computer interaction system, etc. [11].

In order to organize a completely automatic system, classification algorithms are utilized in the combination with a face detection algorithm, which selects candidates for further analysis [12–17]. In this paper we propose a system which extracts all the possible information about depicted people from the input video stream, aggregates and analyses it in order to measure different statistical parameters (Fig. 1).

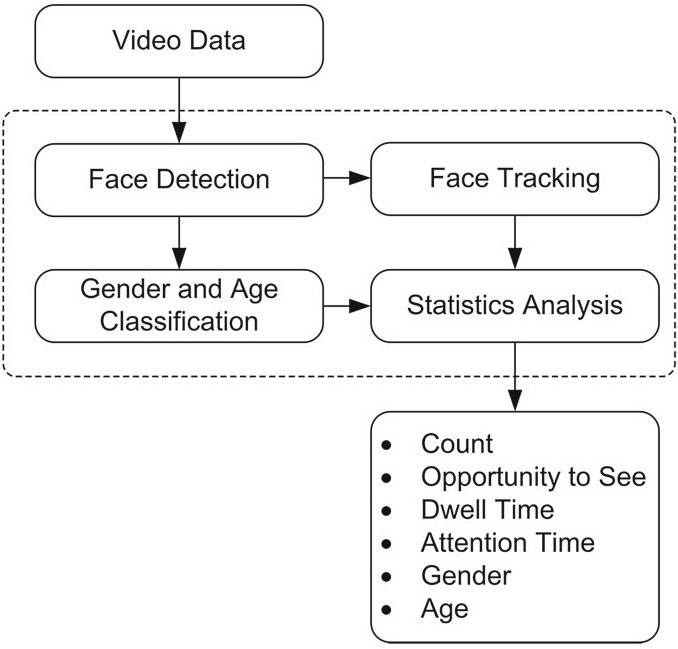



**Fig. 1.** A block diagram of the proposed application for video analysis

The quality of face detection step is critical to the final result of the whole system, as inaccuracies at face position determination can lead to wrong decisions at the stage of recognition. To solve the task of face detection AdaBoost classifier, described in paper [18], is utilized. Detected fragments are preprocessed to align their luminance characteristics and to transform them to uniform scale. On the next stage detected and preprocessed image fragments are passed to the input of gender recognition classifier which makes a decision on their belonging to one of two classes («Male», «Female»). Same fragments are also analyzed by the age estimation algorithm. The proposed gender and age classifiers are based on non-linear SVM (Support Vector Machines) classifier with RBF kernel. To extract information from image fragment and to move to a lower dimension feature space LBP features are utilized.

To estimate the period of a person's stay in the range of camera's visibility, face tracking [19–22] algorithm is used. It is based on Lucas-Kanade optical flow calculation procedure [23].

The rest of the paper briefly describes main algorithmic techniques utilized on the stages of gender and age recognition. The level of gender and age classification accuracy is estimated in real-life situations.

## 2 Gender Recognition

A new gender recognition algorithm, proposed in this paper, is based on non-linear SVM classifier with RBF kernel. Detected fragments are preprocessed to align their luminance characteristics and to transform them to uniform scale. After that to extract information from image fragment and to move to a lower dimension feature space local binary patterns (LBP) [24] operator is utilized. These simple local features have been proved to show good results in application to face recognition tasks. Their calculation procedure is shown in Fig. 2.

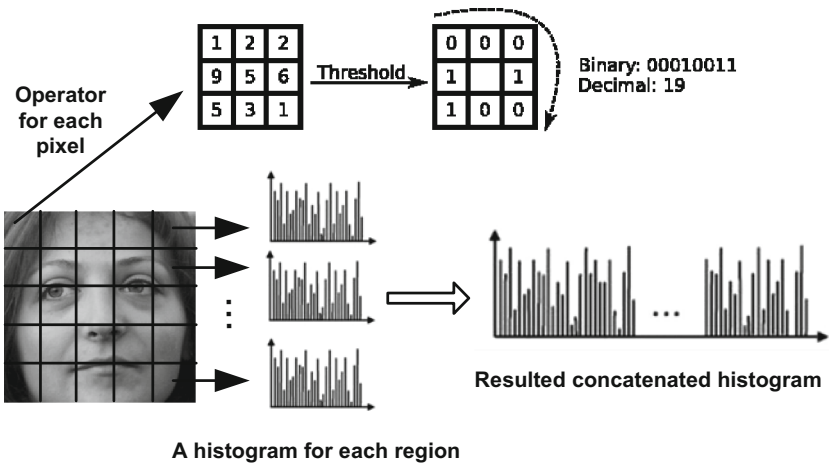



**Fig. 2.** LBP feature vector extraction procedure

On the first step each pixel is compared with its neighbors. The result of comparison is presented in binary scale. These digits from a given neighborhood (let's say 3 × 3 pixels) form a binary number which can be presented in decimal format.

On the second stage image is divided into rectangular regions. A histogram of frequencies of emergence of numbers, acquired on the first step, is calculated for each region. The resulted feature vector is a concatenation of histograms from all regions.

The obtained feature vector is transformed using a Gaussian radial basis function kernel using Eq. 1:

$$k(z_1, z_2) = C \exp\left(\frac{-\|z_1, z_2\|^2}{\sigma^2}\right) \tag{1}$$

Kernel function parameters $C$ and $\sigma$ are defined during training. The resulted feature vector serves as an input to linear SVM classifier which decision rule is specified by Eq. 2:

$$f(AF) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k(X_i, AF) + b\right). \tag{2}$$

The set of support vectors $\{X_i\}$, the sets of coefficients $\{y_i\}$, $\{\alpha_i\}$ and the bias $b$ are obtained at the stage of classifier training. This is how the proposed gender classifier based on LBP features and SVM was constructed (LBP-SVM classifier).

Both gender recognition algorithm training and testing require big enough color image database. The most commonly used image database for the tasks of human faces recognition is the FERET database [25], but it contains insufficient number of faces of different individuals, that's why we collected our own image database, gathered from different sources (Table 1 and Fig. 3).

**Table 1.** The proposed training and testing image database parameters.

| Parameter | Value |
|---|---|
| The total number of images | 8 654 |
| The number of male faces | 5 250 |
| The number of female faces | 5 250 |
| Minimum image resolution | 640 × 480 |
| Color space format | RGB |
| Face position | Frontal |
| People's age | From 18 to 65 years old |
| Race | Caucasian |
| Lighting conditions, background and facial expression | No restrictions |

Faces on the images from the proposed database were detected automatically by AdaBoost face detection algorithm. After that false detections were manually removed, and the resulted dataset consisting 10 500 image fragments (5 250 for each class) was

a) male faces b) female faces

**Fig. 3.** Detected fragments from the proposed image database

obtained. This dataset was split into three independent image sets: training, validation and testing. Training set was utilized for SVM classifier construction. Validation set was required in order to avoid the effect of overtraining during the selection of optimal parameters for the kernel function.

For the representation of classification results we utilized the Receiver Operator Characteristic (ROC-curve). As there are two classes, one of them is considered to be a positive decision and the other – a negative. ROC-curve is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various discrimination threshold settings. The advantage of ROC-curve representation lies in its invariance to the relation between the first and the second error type's costs.

The proposed classifier was compared to AF-SVM algorithm described in paper [10]. AF-SVM was chosen as a reference because it has both high recognition rate and low operational complexity compared to state-of-the-art classifiers [26].

Testing results of the proposed LBP-SVM classifier compared to AF-SVM performance are presented in Table 2 and Fig. 4.

**Table 2.** Recognition rate of LBP-SVM classifier compared to AF-SVM

| Algorithm / Parameter | AF-SVM | | LBP-SVM | |
|---|---|---|---|---|
| Recognition rate | True | False | True | False |
| Classified as "male", % | 90.6 | 9.4 | 90.2 | 9.8 |
| Classified as "female", % | 91 | 9 | 94.3 | 5.7 |
| Total classification rate, % | **90.8** | 9.2 | **92.3** | 7.7 |

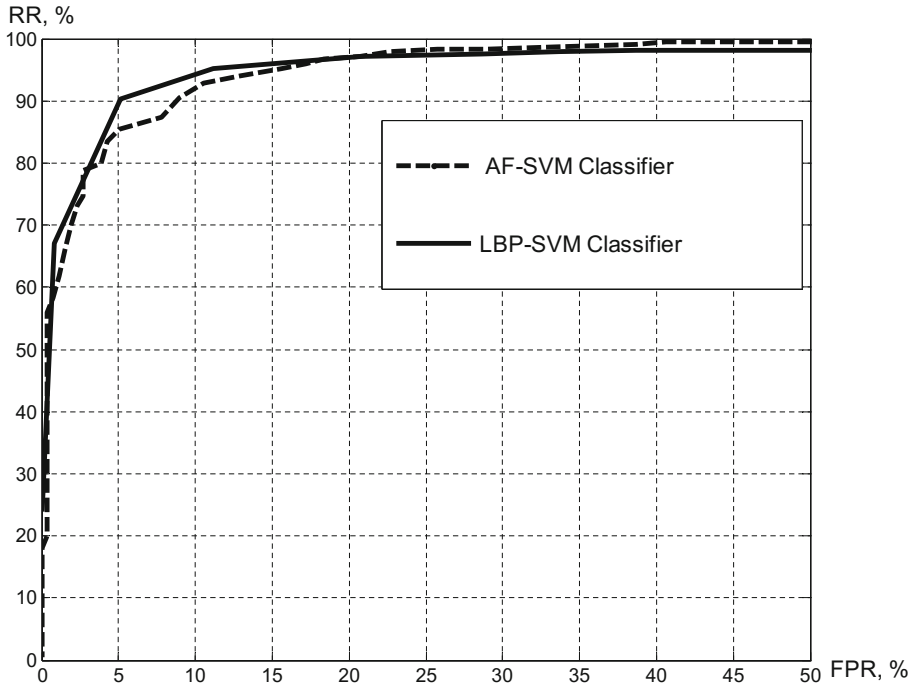



**Fig. 4.** ROC-curves for LBP-SVM and AF-SVM classifiers

Experimental results show that utilization of LBP features for gender recognition improves overall performance by 1.5 % allowing to acquire more than 92 % accuracy.

## 3 Age Estimation

A lot of research in the area of age classification has been done over last few years [27–32]. The proposed age estimation algorithm realizes multiclass classification approach (Fig. 5) where for each age (from 1 to N) a binary classifier is constructed deciding whether a person on input image looks older than the given age or not. Input fragments are preprocessed to align their luminance characteristics and to transform them to uniform scale. Preprocessing includes color space transformation and scaling, both similar to that of gender recognition algorithm. Additionally image normalization was performed by histogram equalization procedure. Transformation to LBP feature space and SVM training procedure are used for binary classifier construction. To predict direct age binary classifier outputs are statistically analyzed and the most probable age becomes the algorithm output.

Training and testing require a huge enough color image database. We used state-of-the-art image databases MORPH [33], FG-NET [34] and our own RUS-FD database of real-life test images which low (60 × 60 pixels on each face) resolution (Table 3). Faces on the images were detected automatically by AdaBoost face detection algorithm.

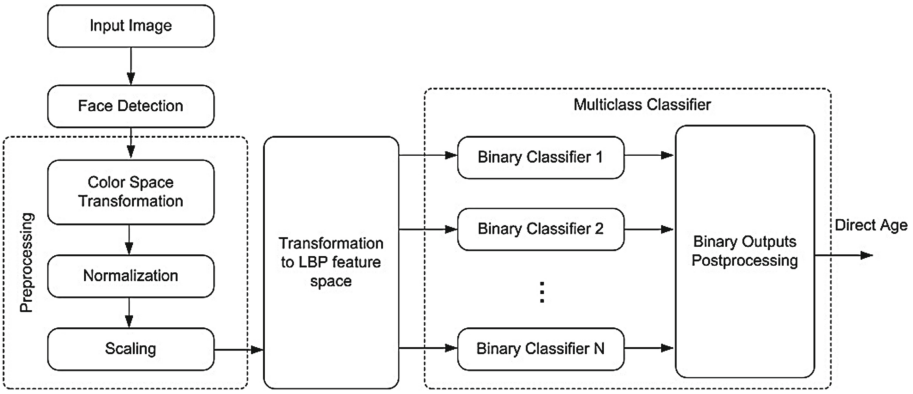



**Fig. 5.** LBP-SVM age estimation algorithm block diagram

**Table 3.** Face databases for age estimation algorithms learning and testing

| Database name | Total number of images using | Range of ages | Age distribution |
|---|---|---|---|
| MORPH | 6513 | 16–75 | Non-uniform |
| FG-NET | 841 | 1–70 | Non-uniform |
| RUS-FD | 6900 | 15–60 | Uniform, 150 faces on each age |

To test age estimation algorithms performance standard metrics were calculated:

- Mean Absolute Error (MAE) – mean absolute difference between estimated and real ages.
- Cumulative Score (CS) – the probability that estimated age lies within an interval $dx$ from real age.
- Probability Density Function of age estimation error.

To estimate the proposed algorithm in real-life situation testing firstly performed on FG-NET database. Age on FG-NET database was marked manually by a group of experts to compare subjective estimation with the algorithm performance. The corresponding dependences for LBP-SVM algorithm simulation are presented in Figs. 6, 7, and 8.

The proposed algorithm shows results comparable to the subjective evaluation in a range of ages from 20 to 35 years. The average absolute error in this range is about 6 years old. Accuracy of LBP-SVM algorithm decreases on senior ages because of MAE grows. In this range (45–60 years), the proposed algorithm yields an expert evaluation approximately 10–15 years in terms of average error.

Cumulative score shows that around 40 % of estimations have less than 5 years deviation from true age and 70 % - less than 10 years deviation. Subjective evaluation curve in Fig. 7 give us the possible limit for future age estimation algorithm improvement.

Analysis of the error probability density function shows that the proposed algorithm has close to symmetric error distribution. Objective results are not inclined to overestimate the true age, which is typical for the evaluation of experts.

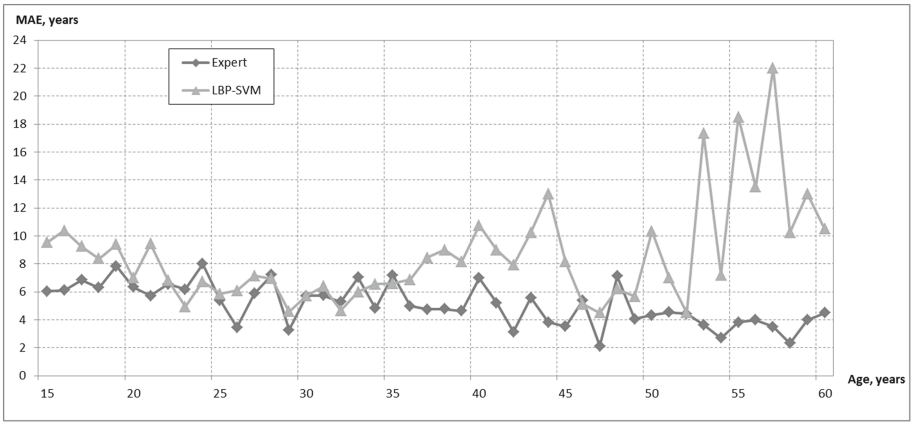



**Fig. 6.** MAE on FG-NET database for the proposed age estimation classifier

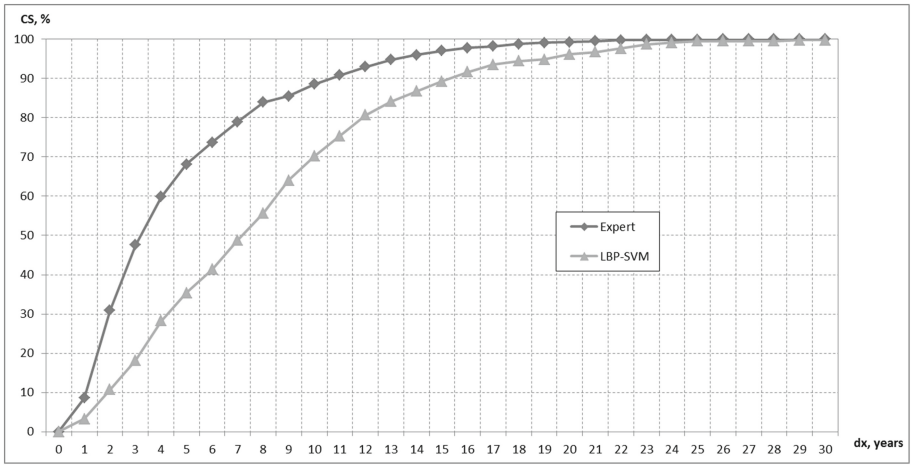



**Fig. 7.** CS on FG-NET database for the proposed age estimation classifier

MAE and CS comparison for LBP-SVM algorithm on different test databases is presented in Fig. 9 and Fig. 10.

Total MAE score of LBP-SVM algorithm on RUS-FD database is 6.94, MORTH database – 7.29, FG-NET database – 7.47. Subjective estimation MAE is 4.2 indicating that the proposed algorithm still needs much improvement to show results comparable to a human. The possible ways to improve the accuracy of age classifier are feature set

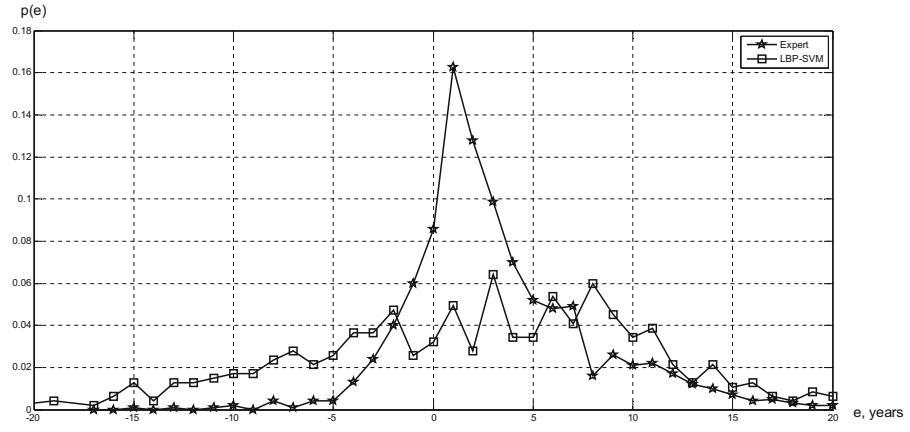



**Fig. 8.** Error probability density function on FG-NET database

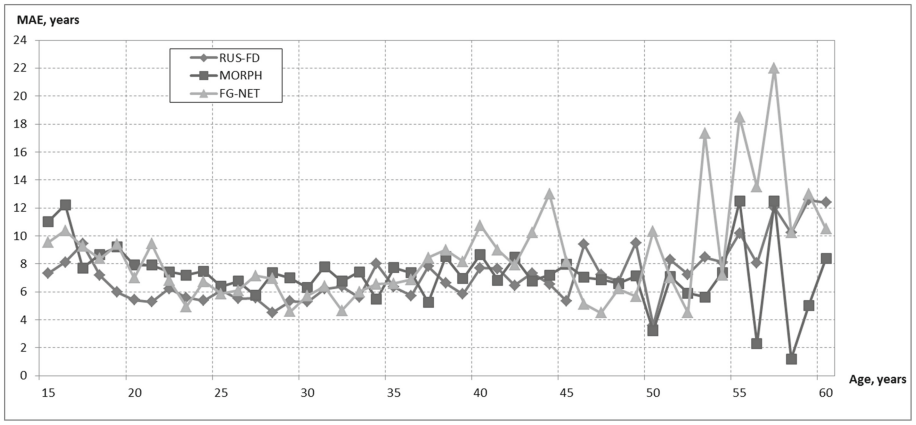



**Fig. 9.** MAE comparison on different databases for LBP-SVM algorithm

expansion (utilization of a combination of different feature transforms), cost-sensitive SVM learning procedure utilization, pre-processing and post-processing steps efficiency improvement.

## 4 Overall Performance Comparison

The proposed audience analysis system is compared to its commercial analog – Intel Audience Impression Metrics Suite (Intel AIM Suite). Experimental setup was the following: an input video stream from IP-camera (Axis M1014) was split into two and analyzed simultaneously by Intel AIM Suite and by the proposed system.

During the experiment a group of people including men and women have been walking in front of the camera imitating difficult situations of movement such as partial

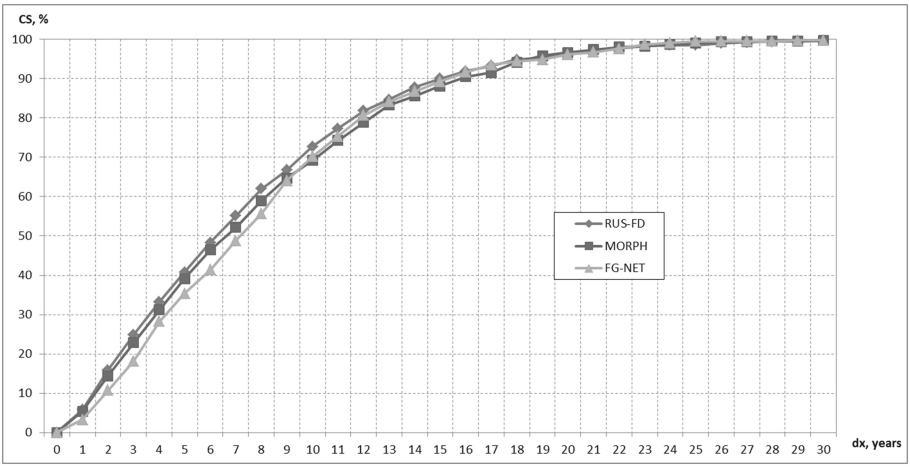



**Fig. 10.** CS comparison on different databases for LBP-SVM algorithm

occlusion and temporary disappearance. The following metric was proposed to compare algorithms performance (Eq. 3):

$$K\frac{D}{N}, \tag{3}$$

where $D$ is the total number of misclassified objects on testing video sequence, and $N$ – the total number of frames. Testing results are presented in Table 4. Experimental results show that Intel AIM Suite seriously overestimates the number of people during people count while the proposed system has higher classification accuracy.

**Table 4.** Audience analysis system comparison results

| System / Parameter | Proposed | Intel AIM Suite | Ground Truth |
|---|---|---|---|
| $K$ | 0.29 | 0.6 | - |
| People count | 16 | 72 | 11 |
| Men count | 15 | 63 | 9 |
| Women count | 1 | 9 | 2 |

## 5 Conclusion

The system, described in this paper, provides collection and processing of information about the audience in real time. It is fully automatic and does not require people to conduct it. No personal information is saved during the process of operation. A modern efficient classification algorithm allows to recognize viewer's gender with more than 92 % accuracy.

The noted features allow applying the proposed system in various spheres of life: places of mass stay of people (stadiums, theaters and shopping centers), transport knots (airports, railway and auto stations), digital signage network optimization, etc.

## References


1. Alpaydin, E.: Introduction to Machine Learning. The MIT Press, Cambridge (2010)
2. Sammut, C., Webb, G.I.: Encyclopedia of Machine Learning. Springer, New York (2011)
3. Li, S.Z., Anil, K.J.: Handbook of Face Recognition. Springer, London (2005)
4. Kriegman, D., Yang, M.H., Ahuja, N.: Detecting faces in images: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **24**(1), 34–58 (2002)
5. Hjelmas, E.: Face detection: a survey. Comput. Vis. Image Underst. **83**(3), 236–274 (2001)
6. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. (CSUR) **35**(4), 399–458 (2003)
7. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, London (2010)
8. Makinen, E., Raisamo, R.: An experimental comparison of gender classification methods. Pattern Recogn. Lett. **29**(10), 1544–1556 (2008)
9. Tamura, S., Kawai, H., Mitsumoto, H.: Male/female identification from 8 to 6 very low resolution face images by neural network. Pattern Recogn. Lett. **29**(2), 331–335 (1996)
10. Khryashchev, V., Priorov, A., Shmaglit, A.L., Golubev, M.: Gender recognition via face area analysis. In: Proceedings of the World Congress on Engineering and Computer Science, Berkeley, USA, pp. 645–649 (2012)
11. Fu, Y., Huang, T.S.: Age synthesis and estimation via faces: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11), 1955–1976 (2010)
12. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 39–51 (1998)
13. Maydt, J., Lienhart, R.: Face detection with support vector machines and a very large set of linear features. In: IEEE ICME 2002, Lousanne, Switzerland (2002)
14. Roth, D., Yang, M.-H., Ahuja, N.: A SNoW-based face detector. In: Solla, S.A., Leen, T.K., Müller, K.-R. (eds.) Advances in Neural Information Processing Systems 12 (NIPS 12), pp. 855–861. MIT Press, Cambridge (2000)
15. Juell, P., Marsh, R.: A hierarchical neural network for human face detection. Pattern Recogn. **29**, 781–787 (1996)
16. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 23–38 (1998)
17. Lin, S.H., Kung, S.Y., Lin, L.J.: Face recognition/detection by probabilistic decision-based neural network. IEEE Trans. Neural Netw. **8**, 114–132 (1997)
18. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 511–518 (2001)
19. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**, 13 (2006)
20. Comaniciu, D., Ramesh, V., Andmeer, P.: Kernel-based object tracking. IEEE Trans. Patt. Analy. Mach. Intell. **25**, 564–575 (2003)
21. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 593–600 (1994)
22. Tao, H., Sawhney, H., Kumar, R.: Object tracking with bayesian estimation of dynamic layer representations. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 75–89 (2002)

23. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of Imaging Understanding Workshop, pp. 121–130 (1981)
24. Da, B., Sang, N.: Local binary pattern based face recognition by estimation of facial distinctive information distribution. Opt. Eng. **48**(11), 117203-1–117203-7 (2009)
25. Phillips, P.J.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
26. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**, 121–167 (1998)
27. Sung, E.C., Youn, J.L., Sung, J.L., Kang, R.P., Jaihie, K.: A comparative study of local feature extraction for age estimation. In: IEEE International Conference on Control Automation Robotics & Vision (ICARCV), pp. 1280–1284 (2010)
28. Thukral, P., Mitra, K., Chellappa, R.: A hierarchical approach for human age estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1529–1532 (2012)
29. Guodong, G., Guowang M.: Human age estimation: What is the influence across race and gender. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 71–78 (2010)
30. Zhen, L., Yun, F., Huang, T.S.: A robust framework for multiview age estimation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9–16 (2010)
31. Guodong, G., Xiaolong, W.: A study on human age estimation under facial expression changes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2547–2553 (2012)
32. Hee, L.W., Jian-Gang, W., Wei-Yun, Y., Xing, L.C., Yap, P.T.: Effects of facial alignment for age estimation. In: IEEE International Conference on Control Automation Robotics & Vision (ICARCV), pp. 644–647 (2010)
33. Ricanek, K., Tesafaye, T.: MORPH: a longitudinal image database of normal adult age-progression. In: IEEE 7th International Conference on Automatic Face and Gesture Recognition, pp. 341–345 (2006)
34. The FG-NET Aging Database. http://www.fgnet.rsunit.com/, http://wwwprima.inrialpes.fr/FGnet/

# Modelling In-Store Consumer Behaviour Using Machine Learning and Digital Signage Audience Measurement Data

Robert Ravnik[1(✉)], Franc Solina[1], and Vesna Zabkar[2]

[1] Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
robert.ravnik@fri.uni-lj.si
[2] Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

**Abstract.** Audience adaptive digital signage is a new emerging technology, where public broadcasting displays adapt their content to the audience demographic and temporal features. The collected audience measurement data can be used as a unique basis for statistical analysis of viewing patterns, interactive display applications and also for further research and observer modelling. Here, we use machine learning methods on real-world digital signage viewership data to predict consumer behaviour in a retail environment, especially oriented towards the purchase decision process and the roles in purchasing situations. A case study is performed on data from a small retail shop where demographic and audience data of 1294 store customers were collected, manually verified and analysed. Among all customers, 246 store customers were involved in a buying process that resulted in an actual purchase. Comparison of different machine learning methods shows that by using support vector machines we can predict with 88.6 % classification accuracy whether a customer will actually make a purchase, which outperforms classification accuracy of a baseline (majority) classifier by 7.5 %. A similar approach can also be used to predict the roles of an individual in the purchase decision process. We show that by extending the audience measurement dataset with additional heuristic features, the support vector machines classifier on average improves the classification accuracy of a baseline classifier by 15 %.

## 1 Introduction

Digital signage systems are nowadays primarily used as public information interfaces. They display general information, advertise content or serve as media for enhanced customer experience [1–4]. The ability to adapt and change broadcasting content in real time -'on the fly'- as well as access to wide audience have made interactive public displays today a highly active and interdisciplinary area of research.

In order to display engaging content and understand interaction of users with digital signage systems, various interaction techniques and case studies


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 123–133, 2014.
DOI: 10.1007/978-3-319-12811-5_9

were performed [5,6]. Different interaction modalities were proposed including hand gestures, gaze, touch, body and face posture [7–10]. Interaction design studies show that the interaction level of users with digital signage systems will increase, including also the mobility of users around the display, if using an unconventional user interface, i.e. the *curiosity object* [11,12]. Müller et al. performed a field study where they observe how passers-by notice and respond to interactivity of digital signage displays. Their observations show that silhouettes which mirror users tend to be the most effective user representation and that it takes time (approximately 1.2s) to notice the interactivity [13]. In parallel to interaction research, audience measurement studies show that also demographic features and varying broadcasting scenarios influence temporal parameters of user attention [14–16].

Since digital signage systems can have a significant effect on commerce, they are also rapidly permeating shopping centers and retail stores. Retail generalization studies reveal that in-store digital signage increases customer traffic and sales [17,18]. Customers seem to be the most responsive to the broadcasting content that relates to the task at hand and their immediate interest. Pantano and Naccarato present retail digital signage systems as an effective way and advantage for retailers to improve the point of sale [19]. Besides being a new communication channel, digital signs present an effective stimulus that improves the image of shopping malls and create a positive influence on consumers shopping experience [20].

In this interdisciplinary paper, we use our custom-developed computer vision enhanced digital signage system capable of collecting audience measurement data to model and predict customer behaviour in an exemplary setup of a small apparel retail store, being able to achieve 88.6 % classification accuracy in predicting the outcome of a purchase process. Using real-world demographic and temporal audience measurement data, which is additionally annotated with retail purchase decision features, we demonstrate a machine learning model that is capable of predicting whether a person is prone to make a purchase or not. The outline of the paper is as follows: Sect. 2 elaborates the purchase decision process, Sect. 3 presents the real-world in-store consumer behaviour dataset, Sect. 4 denotes machine learning results and Sect. 5 provides final conclusions.

## 2 Purchase Decision Process and Digital Signage System for Retail Behaviour

Purchase decision process describes the sequence of actions performed by a customer when deciding to purchase a particular product or service [21]. It can also be described as a process of problem solving, where a consumer satisfies his needs after thoughtful consideration. The outcome of a purchase decision process is a decision whether a customer will buy a given product or service or not.

Purchase decision process can be described with five stages [22]. The first stage is problem recognition where consumer recognizes a problem or a need. The second stage is search for information via heightened attention of consumer

towards information about a certain product, which can even resolve in actual proactive search for information. The third stage represents the evaluation of alternatives, which usually involves a comparison between various options and features based on the models of the expected value and beliefs. In the fourth stage of the purchase decision process a provider, place, time, value, type and quantity of the selected product or service are determined. The fifth and final stage describes the post purchase use, behavior and actions.

We distinguish between three different types of purchase decisions. They differ in value and frequency of purchase, covering different intensity levels of involvement and time invested in the purchase decision [23]: (i) routine response behaviour (for frequently purchased, low involvement products and services), (ii) limited decision-making (unfamiliar brand choices in the known category of products and services), and (iii) extensive decision-making (high involvement, high value and low frequency of purchasing). There are several factors affecting buying behaviour, such as cultural, social and personal decision elements. Cultural factors include cultural context and belonging to a certain social class or subculture. Social factors are defined with position and role of the individual, his family and reference groups, which have a direct or indirect impact on buying behavior. Personal factors are determined with individual's lifestyle, occupation, property status, personality and self-esteem [22,23].

Purchase decision process can involve one or more persons. A set of people that are involved in a single purchase decision is called a buying unit or group. Each member of the group can take up different roles in the purchase decision process [22]:

1. An initiator is the person who recognizes the need and starts with finding the solution by requesting purchase of a product or service. The initiator may be the actual user of the product or he/she could be any other member of the buying group.
2. An influencer is the person whose opinion or position has significant effect on the purchase decision, usually by providing information on product characteristics and evaluation of possible alternatives.
3. A user is the person who will use the product or service. Typically, the user is involved in defining the required product/service characteristics.
4. A decider takes the final decision when choosing between different products. The decision is based on the required characteristics of a product or service.
5. A purchaser has the formal authority to pay for goods or services. Purchaser also determines the terms of purchase, such as the payment method.
6. A passive influencer or a companion is a person who is a member of the purchase group, but is not actively involved in the buying process.

We should comment that the list of roles in the buying group varies throughout the literature; however, the above introduced initiator, influencer, user, decider and purchaser are the key roles, which are used in most of the definitions [22,23]. Note that the role of a passive influencer is not among the commonly defined roles but is introduced in this paper because of the digital

signage audience measurement observations. Passive influencers do not actively participate in purchase process and would otherwise be excluded from further analysis.

A real-world experiment was performed in order to obtain retail audience measurement data. A 24 inch Sony Vaio VPCL135FX/B camera-enhanced computer display was positioned into a small clothing boutique in a city center of Ljubljana (capital of Slovenia, EU). A small retail shop was selected on purpose, to be able to cover the entire retail floor with a single camera unit. The shop's goods were mostly premium priced sports fashion clothing and apparel, which sets the demographic and behaviour characteristics of collected audience measurement data. The experiment was performed within 23 daily sessions, collecting a total of 214 hours of video recordings. Computer display acquired demographic (gender, age group) and temporal features (presence time, in-view time, attention time) of N = 1294 store customers. The experiment was primarily focused in viewership and attention statistics. The analysis reveals that 35 % of visitors *looked-at* the display, having the average attention time of 0.7s. Gender comparison shows that men (1.2s) were more responsive to digital signage than women (0.4s). Significant difference in attention time was also noted when observing age group of observers and broadcasting content. A more detailed description of demographic and temporal audience measurement features as well as results of attention analysis were already published and are available in [15].

For modelling of retail behaviour we relate the collected audience measurement data with additional data on the purchase decision process. The following consumer group features were added: *group-number* which indicates the sequential number of the buying group a person belongs to, *group-size* denotes the total number of people in a given buying group, and a binary parameter *purchase* that describes whether a group made a purchase or not. If the buying group made a purchase, the roles in the purchase decision process of each group member were also denoted, resulting in 6 additional features: *initiator*, *influencer*, *user*, *decider*, *purchaser* and *passive influencer*. The collected data was verified with manual verification of all automatically obtained data (temporal and demographic features) by two human reviewers. They reviewed the recordings and added additional purchase-oriented features to original audience measurement data. In case of disagreement between reviewers, mutual decision was accepted after discussion during the annotation assessment.

## 3 Observed Retail Behaviour Dataset

In our digital signage based experiment of retail behaviour, N = 1294 people visited the store in the time of the experiment, out of which $N_{buy} = 246$ persons were involved in a buying process that resulted in 140 purchases. The distribution of purchase decisions based on demographic features and group size is presented in Table 1. With $n_{all}$ we denote the total number of people whose characteristic meet the given criterion. With $n_{buy}$ we assign the number of persons who meet the selected criterion and have also been involved in the purchase

**Table 1.** Buying process distribution for a given demographic and buying group size feature.

| Feature | Value | $n_{all}$ | $n_{buy}$ | $p_n$ | $p_{buy}$ | $p_{all}$ |
|---|---|---|---|---|---|---|
| Gender | Male | 504 | 92 | 0.39 | 0.37 | **0.18** |
| | Female | 790 | 154 | 0.61 | 0.63 | **0.20** |
| Age group | 1–14 | 95 | 20 | 0.07 | 0.08 | 0.21 |
| | 15–24 | **133** | **12** | **0.10** | **0.05** | **0.09** |
| | 25–34 | 258 | 54 | 0.20 | 0.22 | 0.21 |
| | 35–44 | 323 | 60 | 0.25 | 0.24 | 0.19 |
| | 45–54 | 251 | 53 | 0.19 | 0.22 | 0.21 |
| | 55–64 | 153 | 30 | 0.12 | 0.12 | 0.20 |
| | 65+ | 81 | 17 | 0.06 | 0.07 | 0.21 |
| Group size | 1 | 438 | 57 | 0.34 | 0.23 | **0.13** |
| | 2 | 618 | 124 | 0.48 | 0.50 | **0.20** |
| | 3 | 165 | 57 | **0.13** | **0.23** | **0.35** |
| | 4 | 68 | 8 | 0.05 | 0.03 | 0.12 |
| | 5 | 5 | 0 | 0.004 | 0.0 | 0.0 |
| Overall | | 1294 | 246 | | | |

decision-making process which resulted in a purchase. Probability $p_n$ is defined as the ratio between the occurrence of a given criterion and the total number of participants $p_n = n_{all}/\mathrm{N}$. We also define the occurrence probability within the people who have actually made the purchase process as $p_{buy} = n_{buy}/\mathrm{N_{buy}}$. Considering the distribution within a single feature space we also define normalized probability $p_{all}$ as $p_{all} = n_{buy}/n_{all}$.

Among all shop visitors, there were 61 % female customers ($p_n = 0.61$) which represent 63 % of people making purchase during observed period ($p_{buy} = 0.63$). Men present 39 % of all customers ($p_n = 0.39$) and 37 % of all people making purchase ($p_{buy} = 0.37$). Gender comparison of normalized probability ($p_{all}$) shows that the probability to be involved in the buying process and eventually made the purchase is approximately the same for both men and women (∼20 %).

Age comparison shows certain degree of balance between age groups as almost all normalized probability ($p_{all}$) are around 20 %. The only deviant exception is the age group between 15 and 24 years with $p_{all} = 0.09$.

Group size analysis reveals that during the experiment there were 438 customers who visited the store alone, 618 in buying groups of two, 165 in buying groups of three, 68 in buying groups of four and 5 in group of five customers. We observe an interesting pattern which shows that the size of the buying group importantly affects the probability of purchase. Out of 438 individual customers, only 13 % ($p_{all} = 0.13$) made a purchase decision. 618 persons, representing buying groups of two, have a normalized purchase probability of $p_{all} = 0.20$. The highest purchase probability have three-person groups with $p_{all} = 0.35$.

**Table 2.** The distribution of roles in purchase decision process for a given demographic and buying group size feature.

| Feature | Value | $n_{init}$ | $n_{inf}$ | $n_{dcdr}$ | $n_{prch}$ | $n_{usr}$ | $n_{pssv}$ | $p_{init}$ | $p_{inf}$ | $p_{dcdr}$ | $p_{prch}$ | $p_{usr}$ | $p_{pssv}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | Male | 48 | 7 | 53 | 56 | 57 | 28 | 0.52 | **0.08** | 0.58 | 0.61 | 0.62 | **0.30** |
| | Female | 106 | 60 | 93 | 85 | 76 | 12 | 0.69 | **0.39** | 0.60 | 0.55 | 0.49 | **0.08** |
| Age group | 1–14 | 1 | 0 | 1 | 1 | 1 | 19 | 0.05 | **0.0** | 0.05 | 0.05 | 0.05 | **0.95** |
| | 15–24 | 5 | 2 | 5 | 4 | 5 | 5 | 0.42 | 0.17 | 0.42 | 0.33 | 0.42 | **0.42** |
| | 25–34 | 39 | 18 | 35 | 36 | 33 | 1 | 0.72 | 0.33 | 0.65 | 0.67 | 0.61 | 0.02 |
| | 35–44 | 41 | 18 | 39 | 41 | 33 | 5 | 0.68 | 0.30 | 0.65 | 0.68 | 0.55 | 0.08 |
| | 45–54 | 40 | 16 | 36 | 31 | 32 | 4 | 0.76 | 0.30 | 0.68 | 0.59 | 0.60 | 0.08 |
| | 55–64 | 18 | 9 | 19 | 20 | 18 | 3 | 0.60 | 0.30 | 0.63 | 0.67 | 0.60 | 0.10 |
| | 65+ | 10 | 4 | 11 | 8 | 11 | 3 | 0.59 | 0.24 | 0.65 | 0.47 | 0.65 | 0.18 |
| Group size | 1 | 57 | 2 | 56 | 55 | 50 | 0 | 1.0 | 0.04 | 0.99 | 0.97 | 0.88 | **0.0** |
| | 2 | 73 | 47 | 68 | 64 | 63 | 14 | 0.59 | 0.38 | 0.55 | 0.52 | 0.51 | **0.11** |
| | 3 | 22 | 16 | 20 | 20 | 18 | 23 | 0.39 | 0.28 | 0.35 | 0.35 | 0.32 | **0.40** |
| | 4 | 2 | 2 | 2 | 2 | 2 | 3 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | **0.38** |
| Overall | | 154 | 67 | 146 | 141 | 133 | 40 | 0.63 | **0.27** | 0.59 | 0.57 | 0.54 | **0.16** |

This group size also achieves a high conversion rate between $p_n = 0.13$ and $p_{buy} = 0.23$.

The distribution of roles in a group is the crucial element in the purchasing process. Every member of a buying group took up at least one of the six roles. Several members can take up the same role (e.g., there may be several users or influencers), and also oppositely, one individual can take up multiple roles. Table 2 shows the distribution of roles in the purchasing process according to the their calculated probabilities. Value $n$ denotes the number of customers in a given group and $p$ the probability of occurrence of this group. Index mark *init* refers to the role of initiator, *inf* to influencer, *dcdr* to decider, *prch* to purchaser, *usr* to user, and *pssv* to passive influencer.

The comparison by gender reveals a significant difference for the roles of influencer and passive observer. Only 8 % of all male customers that participated in a completed purchase ($p_{inf} = 0.08$) have taken up the role of an influencer. The probability for a female customer to take up the influencer role is 39 % ($p_{inf} = 0.39$) which is almost five times higher. The opposite observations hold for the role of a passive influencer, which men took up with probability of 30 % ($p_{pssv} = 0.30$) and women with 8 % ($p_{pssv} = 0.08$).

Age group analysis reveals that youngest age group hardly actively participated in the buying process. Customers aged between 1 and 14 years never took the role of influencer ($p_{inf} = 0.00$) and almost always took the role of passive observer ($p_{pssv} = 0.95$). A very likely explanation for this observation could be the retail's assortment targeted for adult customers.

Group size comparison shows increased correlation between the group size and the probability for its members to take up the role of a passive influencer. For the buying group of two, the probability of a group member to be a passive influencer is 11 % ($p_{pssv} = 0.11$). Probability increases to 40 % ($p_{pssv} = 0.40$) for a member in a buying group of three. As expected, the probability of roles:

**Table 3.** Comparison of machine learning algorithms for classification of decision-making process of the purchase.

| Method | CA | Sensitivity | Specificity |
|---|---|---|---|
| Maj | **0.810** | **0.000** | **1.000** |
| NB | 0.867 | 0.768 | 0.890 |
| kNN | 0.851 | 0.492 | 0.935 |
| SVM | **0.886** | **0.594** | **0.954** |
| RF | 0.873 | 0.394 | 0.986 |

initiator, decider, purchaser and user descend inversely linear with the size of the buying group.

## 4 Modelling Retail Behaviour and Purchase Decision Process

The major contribution of this paper is the finding that it is possible to *predict* the purchase decisions and roles in the purchase decision process by using machine learning methods on our digital signage audience measurement data. Audience measurement data which is additionally annotated with purchase decisions and purchase roles is used to train purchase decision classifiers. Several machine learning algorithms were used in order to compare classifiers of retail behaviour. The baseline classification accuracy for our analysis is 81 %, which corresponds to the *apriori* probability of the class distribution as there were 246 (19 %) out of 1294 consumers that made a purchase during the period of the experiment.

Table 3 shows the results of 10-fold cross-validation for different machine learning methods: the majority classifier (Maj), the naive Bayesian classifier (NB), K-Nearest Neighbor (kNN), Support Vector Machines (SVM) and Random forest (RF). Classification accuracy represents the ratio of correctly classified examples. Sensitivity (also true positive rate or recall) denotes the ratio between the correctly classified positive examples and the number of all positive examples in machine learning dataset. Similar measure is specificity (also true negative rate) which measures the ratio between correctly classified negative examples and the number of all negative examples. The target class value is set to whether a person was involved in a purchase (purchase = yes). Majority classifier reaches a 81 % classification accuracy which sets the baseline for comparison with other methods. The best classification results using 10-fold cross validation are obtained with the SVM classifier which reaches a classification accuracy of 88.6 % and improves the prediction of the majority classifier for 7.6 %. SVM also achieves best sensitivity and specificity rates of 59.4 % and 95.4 % respectively. The random forest method turns out to be the second best, reaching 87.3 % classification accuracy. Other methods of machine learning improve the baseline's classification accuracy for ∼5 %.

**Table 4.** Comparison of machine learning algorithms for classification of roles in purchase decision process.

| | Role in the purchase decision process | | | | | |
|---|---|---|---|---|---|---|
| Method | Initiator | Influencer | User | Decider | Purchaser | Passive Infl. |
| Maj | **0.626** | **0.728** | **0.541** | **0.593** | **0.573** | **0.837** |
| NB | 0.679 | 0.735 | 0.606 | 0.614 | 0.614 | 0.882 |
| kNN | 0.659 | 0.740 | 0.630 | 0.603 | 0.651 | 0.865 |
| SVM | **0.692** | **0.748** | **0.728** | 0.599 | **0.724** | **0.910** |
| RF | 0.651 | 0.724 | 0.611 | **0.635** | 0.653 | 0.861 |
| $NB_h$ | 0.728 | 0.736 | 0.658 | 0.719 | 0.682 | 0.857 |
| $kNN_h$ | 0.747 | 0.757 | 0.682 | 0.695 | 0.744 | 0.878 |
| $SVM_h$ | **0.793** | **0.768** | **0.731** | **0.764** | **0.755** | **0.918** |
| $RF_h$ | 0.711 | 0.724 | 0.686 | 0.670 | 0.614 | 0.846 |

The presented results show that the dataset which originated from our digital signage audience measurement data can be used for modelling the retail behaviour. Among all people that entered the store, we can predict whether they will be included in a shopping decision with a 88 % classification accuracy, based on observable demographic characteristics and the size of the buying group. We believe that such results can yield significant difference for retailers.

We use a similar approach to model the roles in the purchase decision process. Based on observations noted during the ground truth annotation of the audience measurement data, additional heuristic attributes are defined and added to the dataset. Heuristic binary feature *in-group* indicates whether a person was shopping alone or was part of a larger group. Based on presence time intervals we define a numeric heuristic feature *entered* which describes the sequence number in which group members entered the store. Additional heuristic binary feature *entered-first* is derived from it. In a similar manner, we construct also heuristic features *left* and *left-last*. A set of new heuristic features conclude the ratio between presence and in-view time, and ratio between in-view time and attention time.

The evaluation of role classifiers is performed by a 10-fold cross validation. Each algorithm is tested twice, once with retail behaviour data and the second time with added heuristic features. The classification accuracy of each algorithm is presented in Table 4. The case where heuristic features were also present in the learning dataset is denoted with index $h$.

Again, the most robust and efficient method is SVM. When evaluated on basic retail behaviour dataset (upper half of Table 4) the SVM achieves the best classification accuracy in predicting roles of: initiator (69.2 %), influencer (74.8 %), user (72.8 %), purchaser (72.4 %) and passive influencer (91.0 %). Random forest achieves best classification accuracy of 63.5 % for the role of decider. The classification accuracy of best methods in average exceeds the baseline (majority) classifier for 8.9 %.

The lower part of Table 4 represents the classification accuracy of the selected methods when evaluated on the dataset with added heuristic features. We observe that by adding heuristic features, the best classification accuracy achieved improves for all 6 roles. The best results are obtained when heuristic features are added and with the SVM classifier which outperforms the baseline classifier by 14.9 % on average.

The described in-store consumer behaviour model is based on audience measurement data closely tied to a specific broadcasting location and time when it was collected. Therefore, these results should be understood in the context of the store where our experiment took place. However, we believe that the proposed approach exposes and quantifies certain relationships and behavioural patterns that were already identified before in consumer behaviour literature. It also enables that the measurement and modelling is performed again at an arbitrary location.

## 5 Conclusion

Audience-aware public displays are currently a hot topic of research. The ability to broadcast interactive and targeted content and to collect demographic data of viewers opens the way for interdisciplinary research and broadens the application options of such advanced digital signage systems. This work introduces a new approach to automatic modelling of in-store consumer behaviour based on audience measurement data. The experimental results show that under controlled environment the viewership data can be used to predict purchase decisions. The same model with additional heuristic features can also be used to predict more distinctive characteristics, such as an individual's role in the purchase decision process. We believe that these interdisciplinary results show that digital signage audience measurement data can be used to model various user behaviour.

The presented results open new exciting routes to explore in-store consumer behaviour modelling in combination with data from other sources, such as a shop's assortment and customer database. A comparison with additional retailing audience measurement experiments could also illuminate interesting marketing and consumer behaviour phenomena. The proposed approach could also be used to model user behaviour in different situations, such as edutainment, interaction user interface design and gaming.

## References

1. Lundström, L.I.: Digital Signage Broadcasting: Content Management and Distribution Techniques. Focal Press Media Technology Professional. Focal Press, Oxford (2008)
2. Krumm, J.: Ubiquitous advertising: the killer application for the 21st century. IEEE Pervasive Comput. **10**(1), 66–73 (2011)
3. Alt, F., Müller, J., Schmidt, A.: Advertising on public display networks. IEEE Comput. **45**(5), 50–56 (2012)

4. Kostakos, V., Ojala, T.: Public displays invade urban spaces. IEEE Pervasive Comput. **12**(1), 8–13 (2013)
5. Kuikkaniemi, K., Jacucci, G., Turpeinen, M., Hoggan, E.E., Müller, J.: From space to stage: how interactive screens will change urban life. IEEE Comput. **44**(6), 40–47 (2011)
6. Boring, S., Baur, D.: Making public displays interactive everywhere. IEEE Comput. Graphics Appl. **33**(2), 28–36 (2013)
7. Chen, Q., Malric, F., Zhang, Y., Abid, M., Cordeiro, A., Petriu, E.M., Georganas, N.D.: Interacting with digital signage using hand gestures. In: Kamel, M., Campilho, A. (eds.) ICIAR 2009. LNCS, vol. 5627, pp. 347–358. Springer, Heidelberg (2009)
8. Müller, J., Alt, F., Michelis, D., Schmidt, A.: Requirements and design space for interactive public displays. In: Bimbo, A.D., Chang, S.F., Smeulders, A.W.M., (eds.) ACM Multimedia, ACM, pp. 1285–1294 (2010)
9. Hachet, M., de la Riviere, J.B., Laviole, J., Cohé, A., Cursan, S.: Touch-based interfaces for interacting with 3d content in public exhibitions. IEEE Comput. Graphics Appl. **33**(2), 80–85 (2013)
10. Ren, G., Li, C., O'Neill, E., Willis, P.: 3d freehand gestural navigation for interactive public displays. IEEE Comput. Graphics Appl. **33**(2), 47–55 (2013)
11. Ojala, T., Kostakos, V., Kukka, H., Heikkinen, T., Lindén, T., Jurmu, M., Hosio, S., Kruger, F., Zanni, D.: Multipurpose interactive public displays in the wild: three years later. IEEE Comput. **45**(5), 42–49 (2012)
12. Houben, S., Weichel, C.: Overcoming interaction blindness through curiosity objects. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, pp. 1539–1544. ACM, New York, NY, USA (2013)
13. Müller, J., Walter, R., Bailly, G., Nischt, M., Alt, F.: Looking glass: a field study on noticing interactivity of a shop window. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, pp. 297–306. ACM, New York, NY, USA (2012)
14. Huang, E.M., Koster, A., Borchers, J.: Overcoming assumptions and uncovering practices: when does the public really look at public displays? In: Indulska, J., Patterson, D.J., Rodden, T., Ott, M. (eds.) PERVASIVE 2008. LNCS, vol. 5013, pp. 228–243. Springer, Heidelberg (2008)
15. Ravnik, R., Solina, F.: Audience measurement of digital signage: quantitative study in real-world environment using computer vision. Interact. Comput. **25**(3), 218–228 (2013)
16. Schmidt, C., Müller, J., Bailly, G.: Screenfinity: Extending the perception area of content on very large public displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pp. 1719–1728. ACM, New York, NY, USA (2013)
17. Burke, R.R.: The third wave of marketing intelligence. In: Krafft, M., Mantrala, M. (eds.) Retailing in the 21st Century: Current and Future Trends, pp. 159–171. Springer, Heidelberg (2006)
18. Burke, R.R.: Behavioral effects of digital signage. J. Advertising Res. **49**(2), 180–185 (2009)
19. Pantano, E., Naccarato, G.: Entertainment in retailing: the influences of advanced technologies. J. Retail. Consum. Serv. **17**(3), 200–204 (2010)
20. Newman, A., Dennis, C., Wright, L.T., Kingh, T.: Shoppers' experiences of digital signage-a cross-national qualitative study. Int. J. Digit. Content Technol. Appl. **4**(7), 50–57 (2010)

21. Nicosia, F.: Consumer Decision Processes: Marketing and Advertising Implications. Behavioral Sciences in Business Series. Prentice-Hall, Englewood Cliffs (1966)
22. Kotler, P., Keller, K.L.: Marketing Management. Pearson Prentice Hall, Upper Saddle River (2006)
23. Solomon, M.: Consumer Behaviour: A European Perspective. Prentice Hall, London (2006)

# Shopper Behaviour Analysis Based on 3D Situation Awareness Information

Satu-Marja Mäkelä, Sari Järvinen, Tommi Keränen, Mikko Lindholm, and Elena Vildjiounaite(✉)

VTT Technical Research Centre of Finland, Kaitoväylä 1, 90570 Oulu, Finland
{Satu-Marja.Makela, Sari.Jarvinen, Tommi.Keranen, Mikko.Lindholm, Elena.Vildjiounaite}@vtt.fi

**Abstract.** The customer behaviour understanding is of major importance to brick and mortar retail struggling to keep their market share and competing with online retail. In this paper, we propose a customer behaviour tracking solution based on 3D data. We can cover large areas using numerous inexpensive networked 3D sensors for monitoring and tracking people and we have adopted an adaptive background model in order to be able to react to changes in the store environment. Experiments with people tracking and analysis of the trajectories in a department store show that use of inexpensive 3D sensors and lightweight computation allows classifying shopping behaviour into three classes (passers-by, decisive customers, exploratory customers) with 80 % accuracy.

**Keywords:** Behaviour analysis · Shopper behaviour · Depth sensor based people tracking · Depth sensor based situation awareness

## 1 Introduction

The retail market is currently going through a major change. The online retail is growing strongly, but simultaneously the traditional brick and mortar retail is not experiencing similar growth. The revenues are shifting slowly to e-commerce and in order to keep up with the competition offline retail needs to react.

In offline retail the customer experience is of major importance and good personalized relationship could be a notable competitive advantage. In order to meet the customer expectations and keep the expenses down the store performance has to be optimized. The exact, real-time information on customer behaviour will make this possible. Currently the retailers gather customer insight by analysing point-of-sale data after purchases are completed but it does not contain behavioural information. Information on customer routes in the premises, dwelling times at specific locations and shopping behaviour are examples of data the offline retailer can combine with point-of-sale data and use to optimize the business.

We aim at developing commercially viable customer behaviour tracking and analysis solutions benefiting from state-of-the-art 3D computer vision technology and in the same time keeping the cost-effectiveness as a high priority. In this paper we present a low-cost low-power depth sensor based people tracking solution and our shopping behaviour classification approach. Experimental results with the data, collected in a


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811 pp. 134–145, 2014.
DOI: 10.1007/978-3-319-12811-5_10

retail environment, demonstrate that shopper behaviour can be fairly accurately classified based on the trajectory information acquired via depth sensors.

The main novelty in our people tracker solution lies in using 3D sensors together with an adaptive background modelling approach. To the best of our knowledge, up to date this approach was employed only for analysis of data from video cameras or for combing colour and depth information. Compared to video cameras, our solution has the following benefits:

- It works well in a wide range of lighting conditions, even in complete darkness.
- Object detection is more straightforward and reliable than with bare color information, and unlike with 2D cameras, shadows cannot be mistaken for objects.
- It is easier to handle occlusions – one object partially or completely covering another – which again has traditionally been a severe problem for 2D cameras.
- Depth sensors are less privacy-threatening as they do not provide actual photographic information.

The rest of the paper is organized as follows: first we present in detail our 3D sensor data based people tracking solution. Then we describe the experimental set-up in a clothing section of a department store, our shopper behaviour classification approach and evaluation results. Last section presents conclusions and discusses future work.

## 2 Related Work

Understanding customer behaviour is an interesting topic for both behavioural scientists and retailers and it has been studied for decades. Up to date various taxonomies of shopping trip types and customer behaviour types were proposed, most well-known being the distinction between utilitarian and hedonic shopping [1]: utilitarian shoppers aim at efficient acquisition of products, while hedonic shoppers may get positive experiences from other shopping aspects than achieving original purchase intent. Retailers are interested in attracting and retaining hedonic consumers because they are fairly likely to engage in non-planned purchases and influence opinions of other shoppers [2]. Retailers are also interested in differentiating between utilitarian customers who need some particular items, and the ones aiming at refreshing their wardrobes. An example of the former is a customer, who needs a turtleneck of a certain colour to match a suit: if she cannot find it in one shop, she would probably to leave for other shops. An example of the latter is a person, who wants to find "something nice and fashionable": she is more likely to be eventually satisfied by offers of a sales assistant, especially if she trusts this particular shop [3].

There are various technology solutions implemented for automatic customer tracking – both commercial solutions and research approaches. Shopping cart tracking can be done by using IP (Internet Protocol) cameras, detecting the IR LED (Infrared Light-Emitting Diode) plates installed on the cart [4], or by using RFID (Radio Frequency Identification). RFID tags can be glued to the shopping baskets and monitored using RFID readers installed in selected locations in a shop [5, 6], or RFID readers can be installed in the shopping carts, and tags distributed [7]. All these solutions provide information on the track of the basket or cart, not corresponding in all cases to the path

of the customer – e.g. the shopping cart can be left for the time needed to fetch a product out of the main path or the same cart can be used by a group of people. One way to track customers is by using personal mobile devices, e.g., smartphones, and WiFi. In [8] a mobile application is provided for the shoppers to interact with the products and the information on shopper activities is recorded for analysis.

Customer behaviour analysis requires more accurate data. Computer vision solutions allow accurate tracking of customers and provide rich information on customer behaviour and activities. An approach based on aerial mounted surveillance cameras is presented in [9]. The customer movements are extracted and analysed at the point-of-sale. The authors have evaluated the feasibility of their approach using data recorded in laboratory environments and introduced applications for retail managers, sales personnel and automated customer services. In [10] analysis of data from video cameras allowed to recognize actions of customers in a shopping mall, such as one-hand waving, picking up items etc. and to infer purchasing intentions. Popa et al. [11] used a combination of fish-eye and high-definition (HD) cameras to monitor different kinds of shopper activities in laboratory and in a shop. Fish eye cameras were used for trajectory analysis and HD - for action recognition. The customer behaviour was classified to goal oriented, disoriented and looking around with accuracy over 90 %.

To the best of our knowledge, up to date use of depth cameras for people tracking in retail environment was not studied. Systems for people tracking in laboratory and home environments employed tracking either from top-view [12–14] or from the oblique angles (different side views) [15]. Our solution employs the top-view approach, where people are subtracted as blobs (objects) from the background and after that the objects are tracked. Usually the background image is generated by averaging method as in [12, 15] or using statistical methods as in [13], whereas adaptive background models were employed either for video data [16] or for combining data from video and depth cameras [17, 18]. In this paper we propose a low cost depth sensor based people tracking system with dynamic background model adaptation to support real world requirements. The system is used for classification of customer behaviour.

## 3 People Tracking Based on 3D Situation Awareness Data

Our people tracking solution is based on 3D monitoring using low-cost depth sensors. The software is designed to support an arbitrary number of depth sensors connected to low-power computation platforms that take care of object detection and feeding data over a network to a server (Fig. 1). The server fuses all connected sensors together into a single global coordinate system and performs unified tracking over the sensor network.

### 3.1 Node Installation and Functionality

Each depth sensor node is calibrated once, generating data on sensor position and angle, as well as on floor location. The floor points are selected manually from the depth sensor view. According to these points a plane describing the floor area is formed

by using minimum least squares method in the distance between the points and the plane. From this plane a transformation matrix is formed describing rotation, scale and translation parameters. This is used to transform 3D points into a millimetre scale overhead view so that floor height is set to zero, positive axis pointing upwards. The algorithm to accomplish this is described in detail in [14]. Then all sensor nodes are calibrated to have the same directional axis.

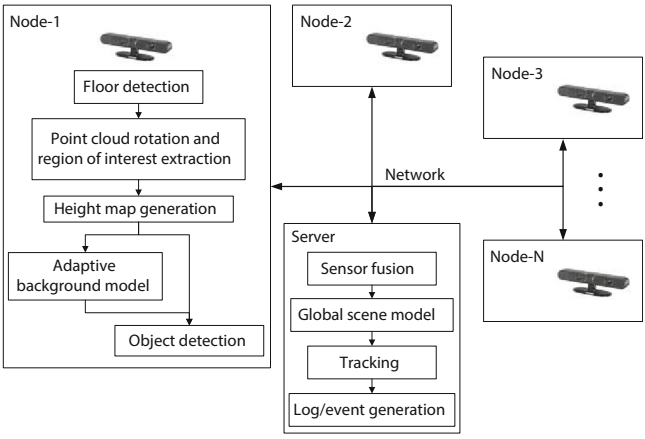



**Fig. 1.** Software structure with the main components highlighted.

In the calibration phase it is also possible to manually modify the sensor view positioning based on the height map visualisation (Fig. 5). For example non-interesting areas can be left out as all the points outside the height map window are ignored.

Adaptive background model is formed in each node. In the start of the detection process the sensor node creates a model of the scene used for object detection (Fig. 2). Several depth frames are converted into 3D point clouds by projecting them from the sensor perspective into a real-world coordinate system. A single depth image results in a sparse representation of the scene filled with holes, but by accumulating several frames, generally 20 depth frames, a more complete model, a topological representation, can be generated. Example of background image can be found in Fig. 5.

To function properly in changing scenes, the background model has to be adapted automatically. Each node has one-frame memory of number of objects detected. If an object was detected in the last frame, but not in the current, a flag is raised and prior to the detection process in the next frame a new background model is created (Fig. 2).

New static objects are recognised and merged into the background model instead of continuing to (falsely) detect them as objects of interest. The tracking server informs the node, when an object is stationary for a specific period. For the next several frames, new information is accumulated in object region, then the background model of this region is updated, and the overall background model is smoothed.

Object detection is also performed in each node by converting each depth frame into a height map and then subtracting the background model from it (Figs. 3 and 5).

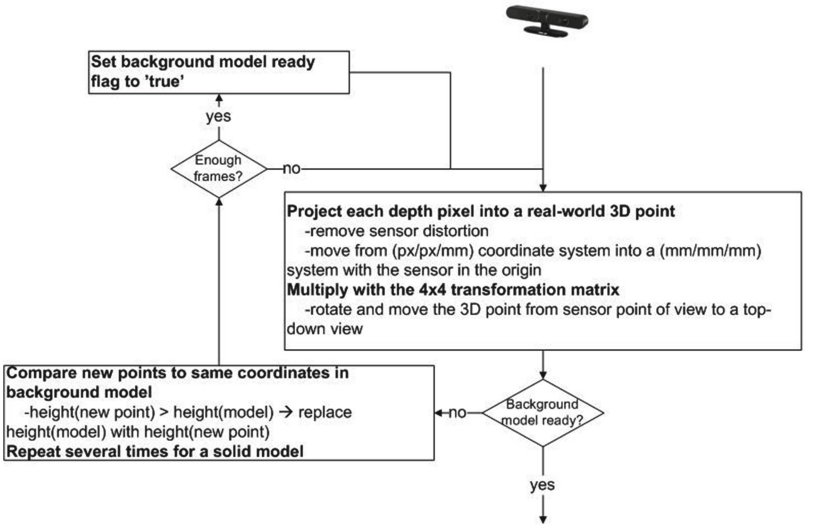



**Fig. 2.** Background model generation process.

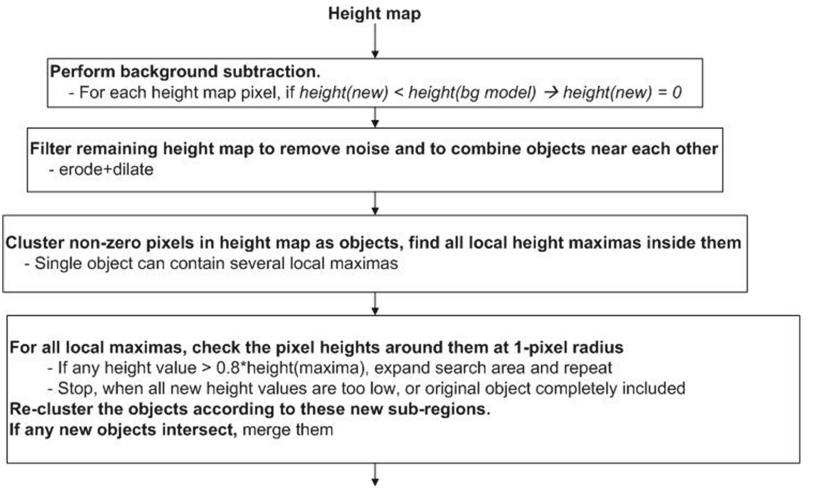



**Fig. 3.** Object detection process.

Clusters, blobs, are extracted from the subtracted image, and then processed by locating their local height maxima, and re-clustering the pixels within the already detected objects so that only pixels close enough in height to the maxima are accepted. This has the effect of dividing some of the larger blobs into several smaller ones, for example to make apart people standing very close to each other. This process also has the tendency

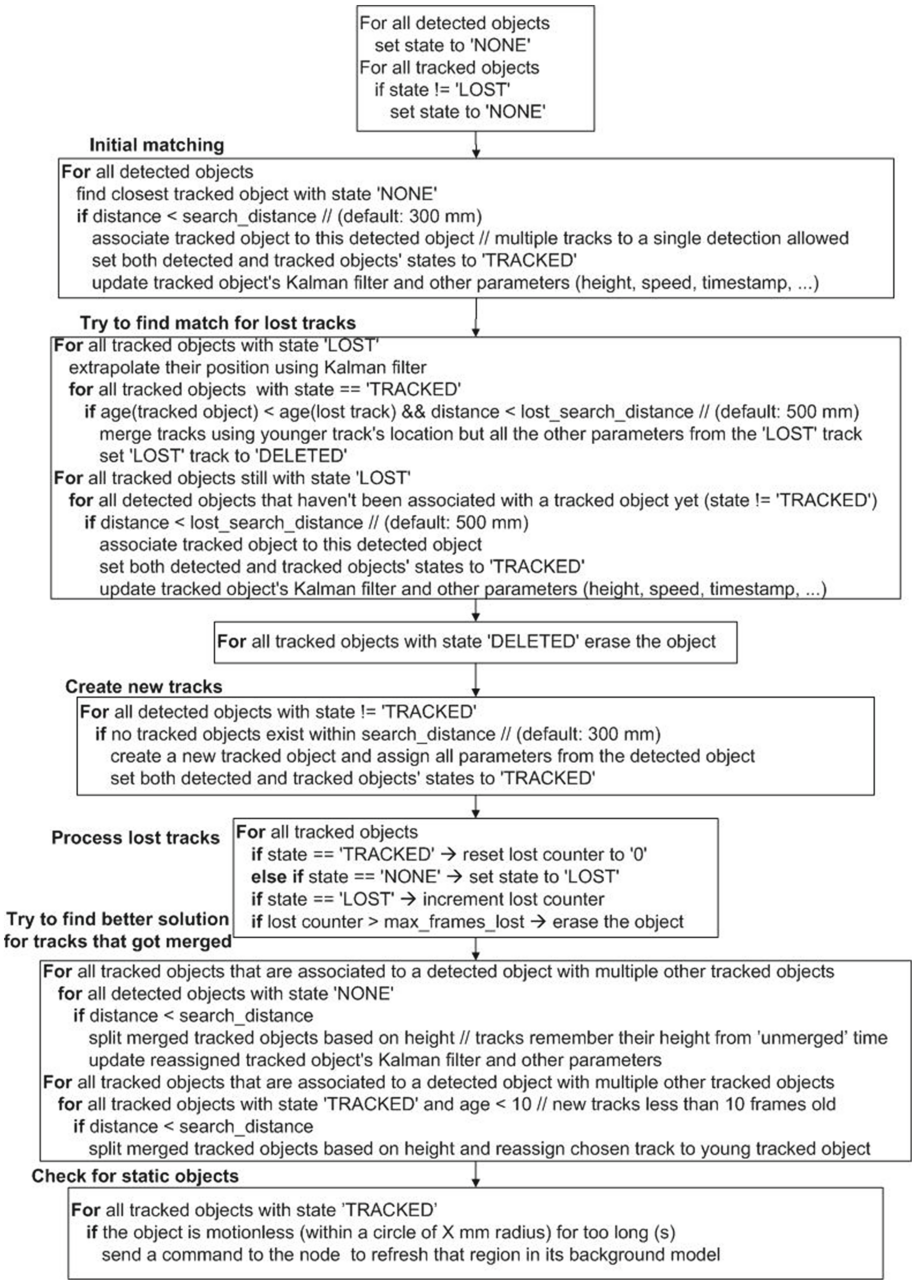



**Fig. 4.** Tracker algorithm process.

to divide a person with an extended arm into several separate objects, but that is something the tracker algorithm on the server side has been designed to deal with, so at this point the false divisions are acceptable.

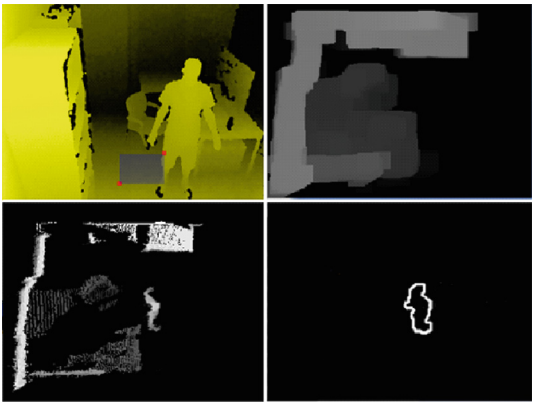


**Fig. 5.** Top-left: depth image with the floor region selected for calibration. Top-right: Background model based on accumulated height maps. Bottom-left: Height map generated from the depth image. Bottom-right: Extracted objects after background subtraction and clustering.

### 3.2 Server Installation and Functionality

All the sensor nodes send their object detection data with unique sensor ID over a network to the server, which combines feeds from number of nodes into one large-scale unified tracker. The server has a visualization window for forming the unified view, on which each sensor node's field of view's height map is presented as a rectangle. The view is adjusted manually to correctly reflect the sensors' real-world positions.

The tracking algorithm on the server uses objects' location, height, age, grouping information and Kalman filter to ensure the best matches from frame to frame and from one sensor field of view to another. The process is described in Fig. 4. The server output includes object ID, position (x, y), height (z), velocity and age in frames for each processed frame and tracked object.

## 4 Test Setup and Data Collection

We set up the depth sensor based people tracker system in a department store in order to collect realistic data for customer behaviour analysis. All together 7 depth sensors were attached to the ceiling with a downward angle and connected to 4 PCs, where the node processes and the server for collecting and combining the tracking data were running. The positioning of the sensors and the covered field of view is depicted in Fig. 6 on the left. The area included main pathways that went through women clothing and cosmetic departments leading to other parts of the department store and also from/to parking space. The people tracker data (position and velocity) were collected from this area in order to recognize different types of customer behaviour. Analysis of shopping trajectories in this work aimed at recognizing three classes of motion patterns:

1. passers-by (monitored area contained a passage to parking lot)
2. decisive customers (this class will be called "quick shoppers" below)
3. exploratory customers (this class will be called "slow shoppers" below).

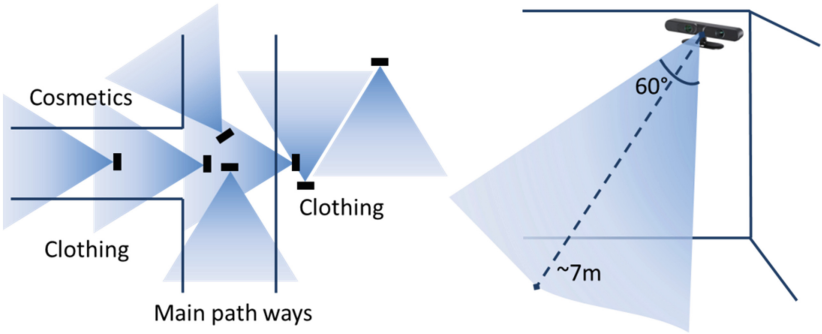



**Fig. 6.** On the left visualization of sensor positions and field of views in our experimental setup and on the right an example of sensor positioning and field of view

The classes were selected because shops need to address them differently: for example exploratory shoppers are most likely ones to need assistance or to check detailed advertisements, whereas other customers may only pay attention to concise advertisements.

We collected data for 35 days, but for the preliminary classification experiments, presented in this paper, we used only 3 days. From collected data we randomly selected one ordinary working day (Tuesday), one Friday and one Saturday. From customer trajectories, acquired on each of these days, we randomly selected 90–100 fairly short trajectories (shorter than 20 s), 120–130 fairly long trajectories (longer than 3 min) and the rest from medium-sized, so that altogether 350 trajectories per day were selected, resulting in a dataset of 1050 customer trajectories.

## 5 Evaluation and Results

### 5.1 Classification Method

Due to the large number of customers in the shops, algorithm to recognize shopping behaviour type should be fairly lightweight. Our classification used individual customer trajectories and their fairly simple characteristics: velocity and points of interests, i.e., small areas where a customer stopped for some time interval. These characteristics were chosen because earlier works observed that consumers, exploring available products, usually move slower in a shop and interact with each product longer than consumers, searching for particular products [19, 20], and that hedonic shoppers tend to spend more time per shopping trip than the ones who do not like shopping as such [2].

From each trajectory the following features were extracted: number of points of interest; maximum time in a point of interest; average velocity during each second and its standard deviation for the whole trajectory. For classification we employed rule-based reasoning, utilizing these features, for example: if maximum time in a point of interest is smaller than threshold T1 and standard deviation of velocity is smaller than threshold T2, then label is “passing by”. Rule parameters, such as thresholds T1 and T2, were learned from the training data by differential evolution algorithm [21],

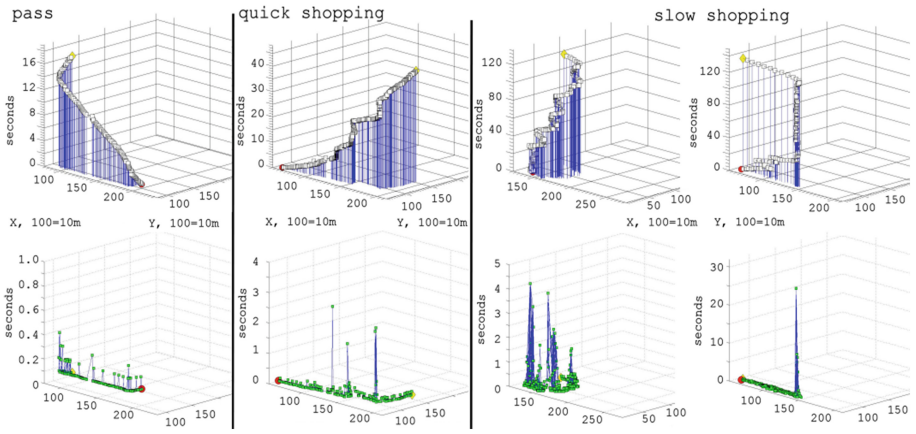



**Fig. 7.** Trajectory examples. Top: 3D trajectory graph (vertical axis shows time); bottom: time spent in each 1 × 1 m area. Left: passing-by, middle: quick shopping; right: two examples of slow shopping.

searching for parameter values that minimize classification error on the training data. Accordingly, rule parameters differed for different training sets and depended on selection of training data.

## 5.2 Data Annotation

The selected dataset was labelled by four annotators (one male and three females), so that each trajectory was annotated by two persons. Annotators had to choose between five labels: "passing-by", "quick shopping: knows what she wants", "slow shopping: search or careful selection", "difficult to say" and "something else". For better understanding how to differentiate between quick and slow shopping, we asked 8 females, aged from 18 to 70 years old, how much time they would need nearby a corresponding stand to fulfil each of the following tasks: (1) to pick a favourite cosmetic product and (2) to find whether a certain blouse or a pullover is not presented (e.g., a blouse of a desired colour and collar type). Time estimates for the first task ranged from 1 to 5 s, and estimates for the second task ranged from 4 to 10 s. Therefore we concluded that a shopping trajectory, containing neither long stops in one area, nor several short stops in a same area, is fairly likely to denote "knows what she wants" behaviour of decisive customers.

During annotation the "something else" label was assigned only in a few cases of short and probably incomplete trajectories, so these cases were excluded from the experiments. From other trajectories we selected the ones that received the same label from two annotators. Only a few cases were labelled as "difficult to say" by two persons, so these cases were also discarded. The resulting dataset contained 798 trajectories. Figure 7 presents visualized trajectories of the three types.

### 5.3 Experimental Protocol and Results

Random selection of trajectories resulted in a fairly imbalanced dataset: it contained a little bit over 100 trajectories for each of the two classes "passers-by" and "quick shoppers", and the rest were "slow shoppers". In training datasets we included equal number of examples from each class, and used remaining data for testing. We experimented with three different sizes of training datasets: 30 samples (10 samples per class), 75 samples (25 samples per class) and 150 samples (50 samples per class). Classification accuracy was evaluated via cross validation, and each training dataset was built by randomly selecting required number of examples per class from samples, not yet used for training. Accordingly, including 50 samples per class in a training set allowed to create two training sets and two test sets; including 25 samples per class in a training set allowed to create four training sets and four test sets; and including 10 samples per class in a training set allowed to create ten training sets and ten test sets. Table 1 presents classification accuracies for different sizes of training datasets, averaged over corresponding numbers of training/test sets.

**Table 1.** Classification accuracies for different sizes of training datasets.

| Num. training samples per class | 10 | | | 25 | | | 50 | | |
|---|---|---|---|---|---|---|---|---|---|
| Overall average accuracy | 74.8 % | | | 79.7 % | | | 81.4 % | | |
| Average accuracy per class | pass | quick | slow | pass | quick | slow | pass | quick | slow |
| | 70 % | 68 % | 77 % | 90 % | 75 % | 79 % | 96 % | 72 % | 81 % |

Table 1 shows that when training data size exceeded 30 samples, "passing-by" behaviour was recognized most accurately, and "quick shopping" – least accurately. The majority of errors were caused by misclassifications of "slow shopping" as "quick shopping", often in cases which were difficult to annotate for humans. For instance, "slow shoppers" walking slowly, but not stopping for a long time, are more difficult to recognize: right "slow shopping" trajectory in Fig. 7 illustrates more prominent example of exploratory shopping than the left one. Part of the errors occurred also due to employing two pairs of annotators, whose views did not match exactly.

The tests demonstrated that 10 training samples per class is too small data size: variations of classification error between different training/test sets were greater than in cases of larger training sets. However, results for training sets, containing 25 and 50 samples per class, are fairly consistent with each other and thus as little as 25 labelled trajectories per class can suffice. Therefore the employed rule-based classification method does not require significant efforts from human annotators and can be easily adapted to specifics of other shops and interests of shop managers: for example, if shop managers would like to draw a line between "quick" and "slow" shoppers differently, or if they would like to detect persons who stayed in one point of interest for a fairly long time, they would only need to label fairly small number of corresponding trajectories and to train new rules.

## 6 Conclusions and Future Work

We have developed a system that can track the customers in real time with sufficient reliability in stores to enable customer behaviour classification. The results show that using low cost depth sensors for monitoring people movement it is possible to obtain information that can be used for detailed customer shopping behaviour analysis. To evaluate the feasibility of our people tracking technology in retail environment we have used the people tracker output to classify the customers based on velocity and points of interests. The classification with simple features performs with high-enough accuracy that it would give meaningful statistical information of customer shopping behaviour to store managers and other stakeholders.

Although the people tracker system setup and sensor positioning was optimised to cover the areas without blind spots, some of those did exist in our system. They were caused by several reasons mainly by low ceiling limiting the sensor view, high and compactly positioned clothing racks and poor overlapping in sensor view. This caused breaking and erroneous merging of customer trajectories. Currently we are modifying the tracker functionalities and post-filtering of the tracking data to improve reconnecting of trajectories.

In future work we will aim for real-time implementation of shopper behaviour classification as our first results suggest that shopping behaviour can be recognized shortly after a customer appears in the tracked area or his behaviour changes. As the customer behaviour and aims often vary during one visit in a store, we plan to develop methods for detecting behaviour change and need in help from sales assistants as early as possible. We also plan to develop more sophisticated algorithms for predicting customer intentions, such as detection of next points of interest.

## References

1. Babin, B.J., Darden, W.R., Griffin, M.: Work and/or fun: measuring hedonic and utilitarian shopping value. J. Consum. Res. **20**(4), 644–656 (1994)
2. Kim, H.-Y., Kim, Y.-K.: Shopping enjoyment and store shopping modes: the moderating influence of chronic time pressure. J. Retail. Consum. Serv. **15**(5), 410–419 (2008)
3. Wesley, S., LeHew, M., Woodside, A.G.: Consumer decision-making styles and mall shopping behavior: building theory using exploratory data analysis and the comparative method. J. Bus. Res. **59**, 535–548 (2006)
4. Celikkan, U., et al.: Capturing supermarket shopper behavior using SmartBasket. In: Digital Information Processing and Communications, pp. 44–53. Springer, Heidelberg (2011)
5. Cabanes, G., Bennani, Y., Dufau-Joel, F.: Mining customers' spatio-temporal behavior data using topographic unsupervised learning. In: International Conference on Machine Learning and Applications, ICMLA '09, pp. 372–377 (2009)
6. Vukovic, M., Lovrek, I., Kraljevic, H.: Discovering shoppers' journey in retail environment by using RFID. Front. Artif. Intell. Appl. **243**, 857–866 (2012)
7. Gomez, J.M.E., Alvarez, A.F.J., Rodriguez, J.B.: Supermarket costumers routes-and-times identifier. In: 2012 IEEE Colombian Communications Conference (COLCOM), pp. 1–5, 16–18 May 2012

8. Ghosh, R., Jain, J., Dekhil, M.: Brickstreams: physical hypermedia driven customer insight. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT '10), pp. 283–284. ACM, New York (2010)
9. Krockel, J., Bodendorf, F.: Customer tracking and tracing data as a basis for service innovations at the point of sale. In: SRII Global Conference (SRII), pp. 691–696 (2012)
10. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 128–135, 29 September – 2 October 2009
11. Popa, M., Rothkrantz, L., Shan, C., Gritti, T., Wiggers, P.: Semantic assessment of shopping behavior using trajectories, shopping related actions, and context information. Pattern Recogn. Lett. **34**(7), 809–819 (2013)
12. Hernandez, D., Castrillon, M., Lorenzo, J.: People counting with re-identification using depth cameras. In: IET Digest 2011, p. 16 (2011)
13. Hansen, D.W., Hansen, M.S., Kirschmeyer, M., Larsen, R., Silvestre, D.: Cluster tracking with Time-of-Flight cameras. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08, pp. 1–6 (2008)
14. Albiol, A., Albiol, A., Oliver, J., Mossi, J.M.: Who is who at different cameras: people re-identification using depth cameras. IET J. Comput. Vis. **6**(5), 378–387 (2012)
15. Han, J., Pauwels, E.J., de Zeeuw, P.M., de With, P.H.N.: Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. IEEE Trans. Consum. Electron. **58**(2), 255–263 (2012)
16. Hofmann, M., Tiefenbacher, P., Rigoll, G.: Background segmentation with feedback: the pixel-based adaptive segmenter. In: Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 38–43 (2012)
17. Fernandez-Sanchez, E.J., Diaz, J., Ros, E.: Background subtraction based on color and depth using active sensors. Sensors **13**(7), 8895–8915 (2013)
18. Ottonelli, S., Spagnolo, P., Mazzeo, P.L.: Improved video segmentation with color and depth using a stereo camera. In: ICIT 2013 (2013)
19. Gil, J., Tobari, E., Lemlij, M., Rose, A., Penn, A.: The differentiating behaviour of shoppers: clustering of individual movement traces in a supermarket. In: 7th International Space Syntax Symposium (2009)
20. Kholod, M., Nakahara, T., Azuma, H., Yada, K.: The influence of shopping path length on purchase behavior in grocery store. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part III. LNCS, vol. 6278, pp. 273–280. Springer, Heidelberg (2010)
21. Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997)

# Shopper Analytics: A Customer Activity Recognition System Using a Distributed RGB-D Camera Network

Daniele Liciotti[1(✉)], Marco Contigiani[1], Emanuele Frontoni[1], Adriano Mancini[1], Primo Zingaretti[1], and Valerio Placidi[2]

[1] Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy
{d.liciotti,m.contigiani,e.frontoni,a.mancini,p.zingaretti}@univpm.it
[2] Grottini Lab srl, Via S. Maria in Potenza, 62017 Porto Recanati, Italy
valerio.placidi@grottinilab.com

**Abstract.** The aim of this paper is to present an integrated system consisted of a RGB-D camera and a software able to monitor shoppers in intelligent retail environments. We propose an innovative low cost smart system that can understand the shoppers' behavior and, in particular, their interactions with the products in the shelves, with the aim to develop an automatic RGB-D technique for video analysis. The system of cameras detects the presence of people and univocally identifies them. Through the depth frames, the system detects the interactions of the shoppers with the products on the shelf and determines if a product is picked up or if the product is taken and then put back and finally, if there is not contact with the products. The system is low cost and easy to install, and experimental results demonstrated that its performances are satisfactory also in real environments.

**Keywords:** Marketing retail · Consumer behavior · Integrated architecture · RGB-D camera · Video analysis

## 1 Introduction

In the last years, the analysis of the human behavior has been of high interest to researchers because its important and different applications, such as: video surveillance [1,2], ambient assisted living [3], analysis of consumer's behavior [4], group interactions and many others. In particular, in the field of intelligent retail environments, numerous studies to investigate how shoppers behave inside a store and how businesses can change strategies to improve sales are emerging. In order to analyse the buyer activity and to solve general aspects of these problems, techniques of artificial intelligence are used and, in particular, vision and image processing. In recent years, the visual analysis of dynamic scenes is one of the most important research activities in computer vision and image understanding [5–7]. When the visual analysis concerns moving scenes, the general


C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 146–157, 2014.
DOI: 10.1007/978-3-319-12811-5_11

method includes following steps: modelling of environments, motion detection, human identification, classification of moving objects, tracking, behavior understanding and data fusion from multiple cameras [7–9]. In this manuscript, we focus the attention on the study of the consumer behavior in a real retail store, in order to recognize human actions [10–13], such as "interacting with the shelf", "picking or releasing a product", "moving in a group", and "knowing most visited areas in the store". Consumers are main actors in the project because the goal is to increase their satisfaction and, therefore, enhance their purchases. Currently, the identification of the shoppers' behavior implements systems of human observation or video recording with traditional cameras. Some tools, such as virtual stores or eye tracking provide incomplete and unrepresentative data because they are based on a small sample of buyers. As a result, by univocally identifying shoppers and automatically analysing their interactions with the products on the shelves and their activities in different zones, our design considerably increases the value of the current marketing research methodologies. Moreover, the main innovation concerns the original use of tracking system, and the other interesting point concerns the real experimental platform described in the results section combined with a vision based statistical approach. Therefore, the project aims to propose an intelligent low-cost embedded system able to univocally identify customers, to analyse behaviors and interactions of shoppers and to provide a large amount of data on which to perform statistics. The automatic extraction of features that univocally recognize each subject in the scene and their movements, provides an important tool to identify important operations concerning marketing strategies. The application implements techniques of image processing such as: background subtraction, low-level segmentation, tracking and finding contours, in order to map a single shopper and/or a group of people within the store that interact with the products on the shelves, defining an ID unique to each visitor filmed by the camera, and classifying these interactions. Paper is organized as following described: Sect. 2 introduces the main aspects of marketing retail and consumer behavior. Section 3 in detail describes the architecture of the system. The experimental setup is described in Sect. 4 and the results are showed in Sect. 5. Last Sect. 6 described conclusions and future works.

## 2 Marketing Retail and Consumer Behavior

The need to associate marketing retail and consumer behavior is born from the necessity to develop theories, strategies and management models compatible with customer behavior. The concept of shop is changed during years becoming not only the place where customers go to buy a specific product, but also the place where the customers go to spend part of their time. Therefore, it is very important to study the consumer behavior so as to investigate the elements of the decision-making process of purchase that determines a particular choice of consumers and how the marketing strategies can influence the customer. Empirical researches on consumer behavior are primarily based on the cognitive approach, which allows to predict and define possible actions that lead to the conclusion

and to suggest implications for communication strategies and marketing. The basic principle of this approach is that individual actions are the result of information processing. The person collects the information, interprets, processes and uses them to take action. Cognitive approaches cannot completely explain the complexity of consumer behavior, which lives in a changing social and cultural context. According to this approach, the choice of purchasing comes from the ability of the products to generate specific sensations, images and emotions. According Perreau [14], five are the steps of consumer buying decision process:

1. *Perception of the problem:* the shopper recognizes a gap between the current situation and desirable situation, therefore perceives a need. The need can be described as a genuine request that comes from the inside and the satisfaction of which is necessary for the survival or to maintain a good level of psychophysical balance.
2. *Research of information:* in order to identify the satisfying solution for the perceived need, the consumer searches for knowledge in the memory, or if the information possessed by the individual is not sufficient, will seek additional data from external sources.
3. *Evaluation of options:* consists of selecting one of the available alternatives based on the criteria defined in the previous step.
4. *Buying decision:* after having identified the place and time.
5. *Post purchase behavior:* is the adequacy of the product purchased and thus the level of consumer satisfaction.

Therefore, the marketing retail discipline defines the set of marketing strategies to point of sale oriented so as to attract the customer and to increase the activities of businesses. To achieve its objectives, the retail marketing uses many techniques through several stages of planning by developing a marketing model for the shop-customer using the most important techniques, in the following described:

- Visual merchandising is the activity of developing floor plans in order to maximize sales. The purpose is to attract, engage and motivate the shopper towards making a purchase. As means of visual merchandising is often widely used a planogram [15].
- Pricing is the activity of establishing the best price that is competitive for shoppers and at the same time with a good profit margin for the store.
- Sensory marketing, to make the shopping experience more pleasant and exciting for the client.
- Loyalty tools, to encourage the consumer to return to the store and to make new purchases.
- Non-conventional marketing concerns original ideas to push the customer to come into the store and trigger a word of mouth process.

The best way to know the behavior of the customer is to create an automatic system that, on the base of acquired knowledge, can predict the purchase of many products and also choices. Therefore, the first goal of this work is to assign an

ID unique to each person detected by a vertical mounted RGB-D camera, to track their activity within the store and then to detect their interactions with the shelf. The next step is to analyse and to classify the interactions: indicating if the product has been picked up and purchased or if the product has been put back after picked up. So, the proposed system will identify the activity of the consumer in front of the shelf.

## 3 Overview of the System Architecture

In order to satisfy both functional and non-functional requirements of the system, a Single Board Computer (for example, Raspberry Pi) has been used, since it is sufficiently small and suited to manage all functions. Functional requirements are: counting and classification of people, their interaction with the shelf, sending data to web server and data analysis; while non-functional requirements are: place of installation and connection modes. As RGB-D sensor, Asus Xtion Pro live has been chosen due to its smaller dimensions than Microsoft Kinect, and the power supply is provided only by USB port. It does not need an additional power.

Figure 1 shows the general scheme of the implemented system and the interactions between the components. The system consists of six devices, listed below:

1. Single Board Computer: is a complete computer built on a single circuit board, with microprocessor(s), memory, input/output (I/O) and other features required of a functional computer. Single-board computers were made as demonstration or development systems, for educational systems, or for use as embedded computer controllers.
2. Asus Xtion Pro live: is composed by an infrared sensor, a RGB sensor and 2 microphones. It is able to provide in output a RGB representation of the scene and also allows to reconstruct a depth map of the same. In the depth map the value of each pixel codifies the distance of each element from 3D scene.
3. Wireless Adaptator.
4. SD/MicroSD Memory Card 8GB Speed 10.
5. Hub USB 2.0: has the task of ensuring the supply of the RGB-D sensor.
6. Router 3G/4G Wireless.

The Single Board Computer uses a SD memory card where Debian operating system is installed allowing an easy configuration of RGB-D sensor of Asus Xtion Pro Live compiling following modules: OpenNI Library[1] and PrimeSense Sensor Driver[2].

The RGB-D sensor is installed (Fig. 2) in a top view configuration at three meters of height from the floor. It visualizes a maximum area (shopper tracking area) of 1.8 m x 3.2 m, but the shelf area (shelf tracking area), that has a height of

[1] https://github.com/OpenNI/
[2] https://github.com/PrimeSense/Sensor

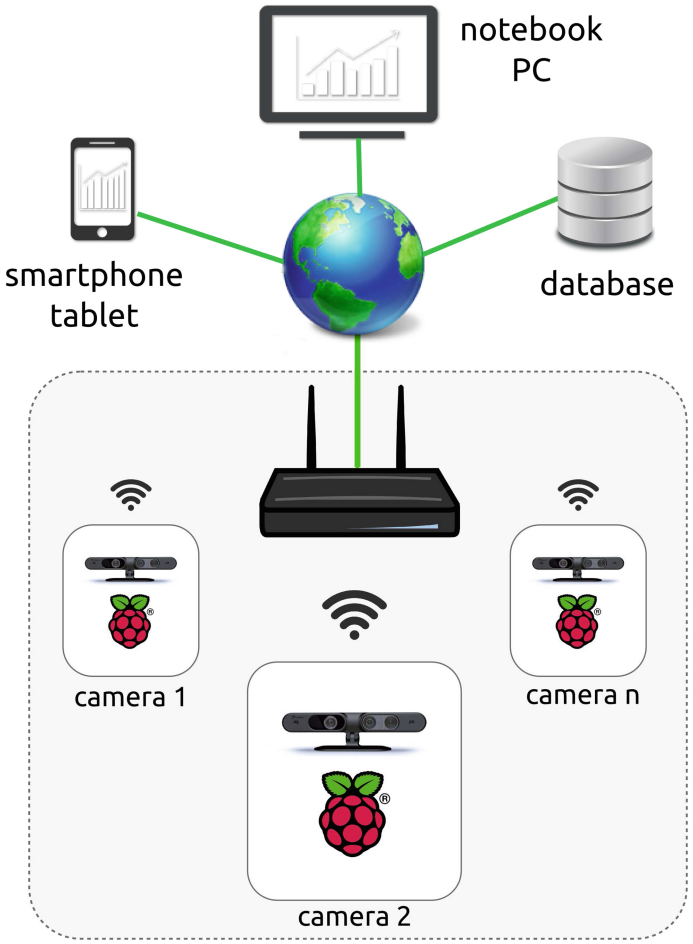



**Fig. 1.** General scheme of connection between the components.

two meters, results smaller than this. The system implements the algorithm that calculates the interactions map between the people in the store and the shelf, sending successively data to a database. Trough a PC, it is possible to connect a smartphone to the database and to visualize the state of system and other interesting information. Figure 3 represents the block diagram that identifies the main steps of the algorithm. The input is the image detected by the camera and the output is the typology of interaction between the user and the products on the shelf.

In the first step, the system acquired the streaming video from the RGB-D sensor. After this, the background subtraction method is implemented, that is one of the most commonly used algorithms for detection of moving objects within a sequence of images. This approach is reliable since each pixel also maintains

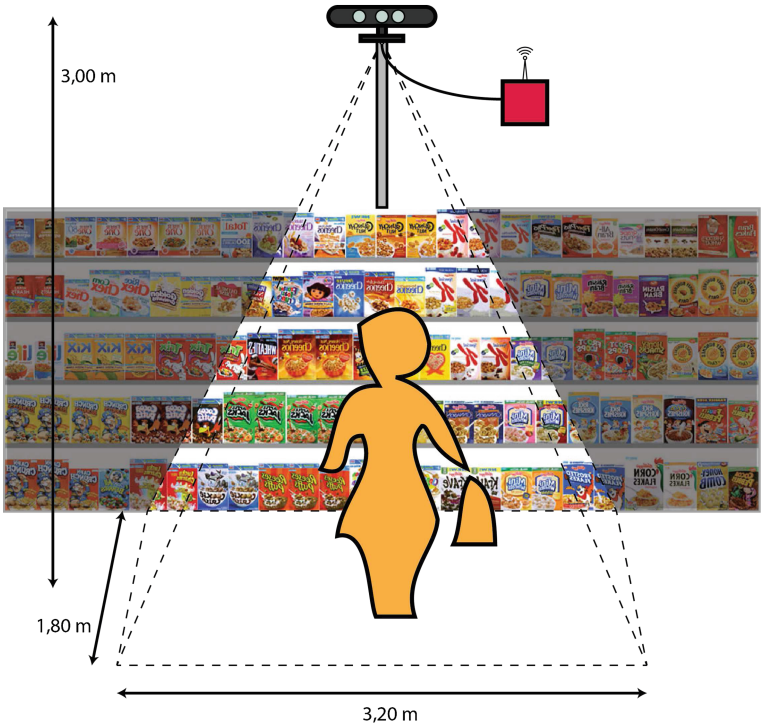



**Fig. 2.** Environment of system installation.

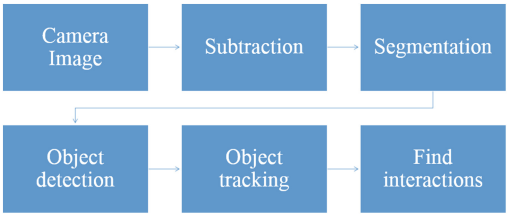



**Fig. 3.** Logical steps of the implemented algorithm.

the depth information, that is not available with a RGB image and so it allows to detect the distance of each blob. Moreover, in order to avoid false detection of objects (false positives), the background image is dynamically updated. After the background subtraction, a threshold value is defined that allows to discriminate positive signals that indicate moving objects, by false positives due to background noise, this method is called segmentation. Another important step consists of the object detection where, for each significant blob, the boundary and the maxima points are found, corresponding to the head of the person. If these points are surrounded by a region of the lowest points comparable to

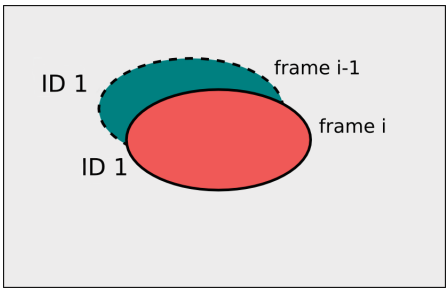



**Fig. 4.** Object tracking implementation.

jump head-shoulder of a human then is a valid blob [16]. The next phase is the object tracking that recognizes the pathways of different blobs along the frames. In other words, in this phase, each blob is recognized and tracked within the streaming video. For each person, the height is determined verifying that this is in the neighborhood of the height of the person in the previous frame. This method is easy but very effective since it is based on the depth image; moreover it is not subject to rapid changes in the forms, allowing a good and reliable tracking. Figure 4 shows how the people are tracked between two successive frames (frame *i-1* and frame *i*). In both frames, the same identifier ($ID_1$) detects the same blob, tracked between frames, so each identifier univocally identifies a person. In this phase of the work, users are not tracked across the sensors, but we retain that this approach must be investigated in future, so that to each visitors maintains a ID unique during the entire visit to the store.

The last step of the algorithm provides the find interactions procedure. When a person has a contact with the shelf, the associate blob is inside the shelf zone. Then, it is possible to detect the exact point of contact by means the definition of common 3-dimensional (XYZ) system coordinates.

The shelf zone, that is defined by user in a configuration file, is formed by three parameters (x shelf dist sx, x shelf dist dx and y shelf dist) as also showed in the following Fig. 5.

When the people interact with the shelf can be presented three different situations, classified as follows:

1. *Positive:* when the product is picked up from the shelf.
2. *Negative:* when the product is taken and then repositioned on the shelf.
3. *Neutral:* if the hand exceeds the threshold without taking anything.

The template matching method has been used to identify and to classify the interactions between the people and the shelf. So, when there is the first contact, the position of the hand in the RGB image is saved, and the same operation occurs when the interaction ends, in order to compare the first image and the final image. If there is a significant correspondence, the interaction is neutral, since there is not an important difference between the first and final image.


**Fig. 5.** Setting parameters of the shelf zone.

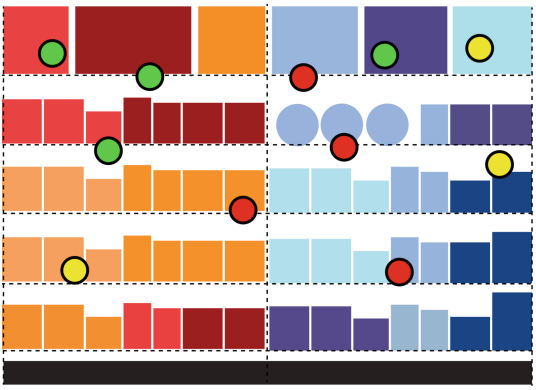


**Fig. 6.** Interactions Map produced by the software (Color figure online).

Otherwise, the interaction can be positive or negative. To identify the type of interaction, the area of the blobs, that is present in the contours, between the two images has been considered.

### 3.1 Interactions Map

This architecture implements a function that displays an interactions map on the screen, during the execution of the program. This is very important because in real time is possible to view the information in which area of the shelf there have been contacts. Figure 6 shows an example of the interactions map. This function draws a colored ball corresponding to the point of contact, on a planogram previously loaded by the user. The color of the ball depends on the type of interaction (green = positive, red = negative and yellow = neutral). This function is used during the debugging phase, in fact all data are saved in a database for later analysis.

## 4 Experimental Setup

The development of the procedure to verify the performances of the system has been realized thanks to the collaboration of Grottini Lab that provided the material and, moreover, allowed to test the system in their laboratory and successively in a real store, partner of Grottini Lab. The collaboration with the partner has been very useful to decide the arrangement of the system, according to functional strategic locations for sales and for the input monitoring. All the system has been installed on a panel in the suspended ceiling of the store. Each system gives in output a significant amount of data that are stored in a database, so that they can be successively analyzed to extract indicators. The final test in the real store has been realized installing four RGB-D cameras for a time period of three months, in order to obtain significant and real data. The cameras monitored the entrance (camera 1), the bleach zone (camera 2), the perfumes zone (camera 3) and the shampoo zone (camera 4). The choice to put a camera near the entrance allowed to exactly count the number of people who entered the store. The indicators that are useful to evaluate the shopper behavior and that can help the store staff to understand their preferences and finally, to increase the sales, are:

- Total number of visitors;
- Total number of shoppers;
- Number of visitors in a particular zone;
- Number of visitors interacting with the shelf;
- Number of interactions for each person;
- Number of visitors becoming shoppers (sales conversion);
- Average visit time.

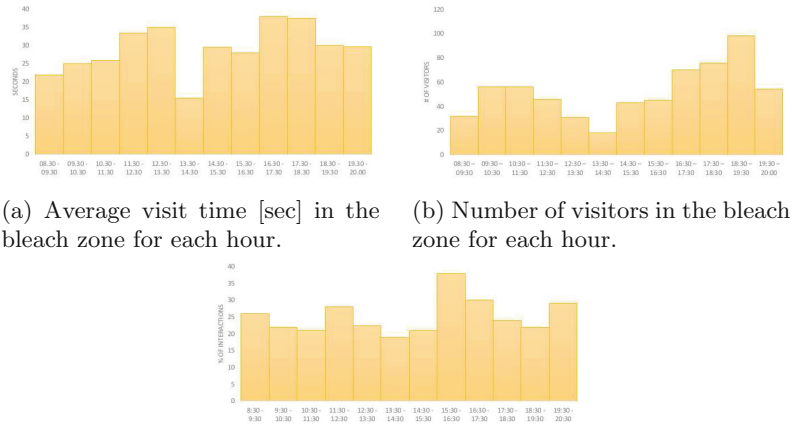
Some indicators that consider the interactions can be:

- Number of products picked up;
- Number of products relocated on the shelf;
- Number of products touched;
- Duration of interactions;
- Average interaction time;
- Number of interactions for product and for category.

## 5 Results

This section presents some experimental results aimed at highlighting the performances of the system in a real environment. The system extracts a high number of parameters, we retain that the most significance to show how it behaves in a real situation are presented in the following graphs.

In the three figures we showed the average trend of some important parameters referred to a day of a week. Figure 7a shows the trend of the average time in



(a) Average visit time [sec] in the bleach zone for each hour.

(b) Number of visitors in the bleach zone for each hour.

(c) Percentage of interactions in the bleach zone for each hour.

**Fig. 7.** Some graphics used to evaluate the shopper behavior.

seconds of visitors in the bleach zone. Observing the graph, the number of visitors is low during the early hours of day (from 8:30 to 9:30) and also during the central hours of day (from 12:30 to 14:30) probably due to less time for making purchases because closer to working time. The other graph that highlights an interesting data is showed in Fig. 7b, which it indicates the trend of the number of visitors always in the bleach zone. From 12:30 to 13:30, in that area, the visitors are 20 on average. This data implies a review of sales strategies, through alternatives marketing solutions as, for example, a more pronounced sign or a temporary closure during the lunchtime. The last Fig. 7c indicates the percentage of visitors that have had at least one interaction with the shelf where the bleach is located. Visitors that have picked up a product or have had a contact with the shelf have been 27 % on average. In order to evaluate the performance of the integrated system that we realized, we compared the number of interactions detected by the system and the real interactions physically determined, by obtaining a reliability factor near to 96 %. The system quite correctly recognize if a visitor crosses along the zone without any stop in front of the shelf, or if he interacts with it. Moreover, the interactions map gives additional visual information on the typology of action that the visitor performed becoming a potential buyer.

## 6 Conclusion and Future Works

In this work, the goal is to propose an automatic and intelligent system able to analyse and to classify the behavior of customers within a retail store. This system received many interest of retailers because, since from the extracted data, they can derive useful information about the behavior of the customer in front

of a shelf. With this technology, it is possible to measure which area of the shelf attracts the attention of the customer, in which shelf to place the products in launch phase, how long the customer remains opposite the shelf and which areas are most visited. From the indicators extracted from the system the retailers can employ a number of marketing strategies in order to attract the customers attention. In the experimental phase, the test of the system in a real environment has provided very interesting results. In particular the architecture is stable, easy to install and especially convenient due to the low cost components. The data provided by the system, that are stored in a database, are very reliable and responsive to the real situation. In fact, in the analysing phase the data, when the interaction is positive, it is possible to correctly identify the product that the customer has picked up. The system ensures a rather high reliability, especially in an ideal condition in which the shelves are those considered to arm height. In the future, the accuracy of the system will be improved independently by the position of the product on the shelf and the consumer position. In addition, the optimization of the image processing algorithm is required, in order to implement an effective procedure of detection and tracking of the shoppers in different areas of the store.

## References

1. Ko, T., Raytheon, C., Arlington, V.A.: A survey on behavior analysis in video surveillance for homeland security applications. In: 37th IEEE Applied Imagery Pattern Recognition Workshop, AIPR '08, pp. 1–8, 15–17 Oct 2008
2. Cristani, M., Raghavendra, R., Del Bue, A., Murino, V.: Human behavior analysis in video surveillance: A Social Signal Processing perspective. Neurocomputing **100**, 86–97 (2013). (Special issue: Behaviours in video)
3. Frontoni, E., Mancini, A., Zingaretti, P.: RGBD Sensors for human activity detection in AAL environments. In: Longhi, S., Siciliano, P., Germani, M., Monteriú, A. (eds.) Living Italian Forum 2013, 300 p. 50 illus, Due: 31 July 2014. Available Formats: eBook ISBN 978-3-319-01118-9
4. Frontoni, E., Raspa, P., Mancini, A., Zingaretti, P., Placidi, V.: Customers' activity recognition in intelligent retail environments. In: Petrosino, A., Maddalena, L., Pala, P. (eds.) ICIAP 2013. LNCS, vol. 8158, pp. 509–516. Springer, Heidelberg (2013)
5. Ascani, A., Frontoni, E., Mancini, A., Zingaretti, P.: Feature group matching for appearance-based localization. In: IEEE/RSJ 2008 International Conference on Intelligent RObots and Systems, IROS 2008, Nice (2008)
6. Ferrari, V., Marin-Jimene, M., Zisserman, A.: Pose search: retrieving people using their pose. In: International Conference on Computer Vision and Pattern Recognition IEEE/CVPR, pp. 1–8 (2009)
7. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: Computer Vision and Pattern Recognition Workshops (IEEE/CVPRW), pp. 9–16 (2010)
8. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: International Conference on Computer Vision and Pattern recognition IEEE/CVPR, pp. 2225–2232 (2011)

9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: International Computer on Vision and Pattern Recognition, IEEE/CVPR, pp. 1778–1785 (2009)
10. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: International Conference on Computer Vision IEEE/ICCV, pp. 1365–1372 (2009)
11. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
12. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
13. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1775–1789 (2009)
14. Perreau, F.: The forces that drive consumer behavior and how to learn from it to increase your sales (2013). theconsumerfactor.com
15. Mankodiya, K., Gandhi, R., Narasimhan, P.: Challenges and opportunities for embedded computing in retail environments. In: Martins, F., Lopes, L., Paulino, H. (eds.) S-CUBE 2012. LNICST, vol. 102, pp. 121–136. Springer, Heidelberg (2012)
16. Migniot, C., Ababsa, F.: 3D human tracking from depth cue in a buying behavior analysis context. In: Wilson, R., Hancock, E., Bors, A., Smith, W. (eds.) CAIP 2013, Part I. LNCS, vol. 8047, pp. 482–489. Springer, Heidelberg (2013)

# Author Index