MNEMOSYNE: Enhancing the Museum Experience through Interactive Media and Visual Profiling

Andrew D. Bagdanov, Alberto Del Bimbo, Lea Landucci, and Federico Pernici

University of Florence,
Media Integration and Communication Center (MICC)
Florence, Italy
{bagdanov,delbimbo,landucci,pernici}@dsi.unifi.it
http://www.micc.unifi.it/

Abstract. MNEMOSYNE is a three year project whose primary goal is to deliver a personalized, interactive multimedia experience to museum visitors through the novel application of personalization driven by computer vision-based profiling. A combination of passive, wall-mounted cameras and sensors carried by guests acquiring active and passive imagery will be used to create a general profile of a museum visitor's interests in order to customize the presentation at interactive tabletop surfaces placed in the museum environment. In this article we discuss the general context in which MNEMOSYNE is defined, as well as the main technical directions the project will follow over the next three years. Some very preliminary results are given for the vision-based techniques to be used for visual profiling of museum visitors.

Keywords: Multimedia interactive museums, personalization, natural interaction, computer vision.

1 Introduction

Museums are traditionally spaces which have, by their very nature, an abundance of information available for dissemination to visitors. This information, however, is also traditionally extremely expensive to place at the disposition of museum visitors due to a number of factors ranging from need for highly qualified curators and exhibit designers to the need to safeguard one-of-a-kind pieces. Because of this, visitor access to museum collections can be limited and at times awkward.

The museum experience can be greatly improved by applying modern techniques of multimedia organization, presentation and interaction and offering different interaction modalities to visitors [24]. Doing this effectively requires a reorganization of the physical and informational space of the museum. A museum must be transformed into an intelligent information space which is, in some sense, aware of the behavior and desires of its visitors, and is able to subsequently provide modes of interaction appropriate to each user. We must in some way unify the physical and virtual museum experiences.

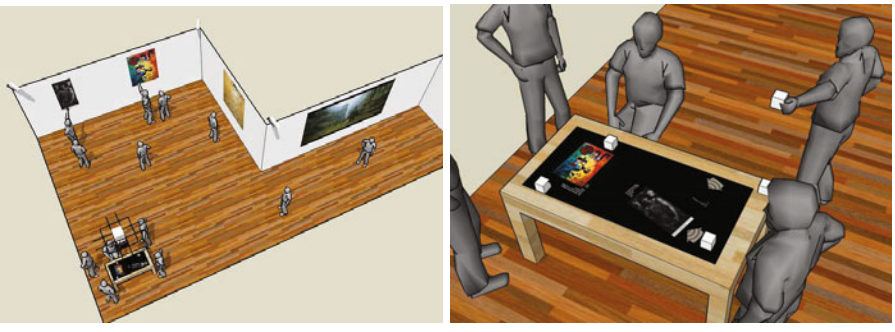
C. Grana and R. Cucchiara (Eds.): MM4CH 2011, CCIS 247, pp. 39–50, 2012.

Access to multimedia information about exhibits can be made highly accessible to non-expert users through the technology of natural interaction [5,6]. However, without appropriate *personalization* of such multimedia information displays, these multimedia stations become little more than fancy Internet kiosks that visitors are uncertain how to access in order to gain more information about aspects of the exhibit of interest to them. Personalization of multimedia museum content is one answer to this problem [1,17]. Personalization offers visitors a customized presentation of appropriate information related to the visitor's tastes and preferences.

In order for personalization to be effective, however, an accurate *profile* of each visitor is necessary, and these profiles should be collected as passively and non-intrusively as possible. Some early approaches to user profiling used sensors attached to, worn or carried by museum visitors. These approaches used first-generation tools and sensors which were very intrusive, such as wearable devices [21]. One of the first attempts was the *Museum Wearable* [20], a wearable computer which orchestrates an audiovisual narration as a function of the visitors interests gathered from his/her physical path in the museum and length of stops. The museum wearable was made by a mobile PC hosted inside a shoulder pack which users had to carry around the museum and a private-eye display that they must wear. Since, according to Donald Norman, the best technology is the one that you just cant see, so easy to use that is becomes "transparent" to the user [16], new less-intrusive solutions have been exploited in recent years. Computer vision technologies have multiple advantages:

- **They are non-intrusive and seamless**
  They are not seen by users, they integrate seamlessly with the architecture, and in many scenarios existing video surveillance infrastructure can be exploited;
- **They are highly scalable**
  The size of deployment can be personal or very large, they can cover a single room of a museum interior, an entire exhibit, or even network multiple museums and multiple exhibits; and
- **They are evolutive and future proof**
  As they rely on cameras, new features and capabilities usually imply software upgrades only.

In a nutshell, the goal of the MNEMOSYNE project is to create an intelligent visual information system capable of constructing a "visual profile" of museum visitors in order to customize interactive multimedia information displays. MNEMOSYNE is a three-year project funded by the Region of Tuscany and the European Commission whose primary goal is to research and develop techniques for delivering a personalized, interactive multimedia experience to museum visitors. The project is challenging as it brings together a number of state-of-the-art and emerging technologies in one application domain. The first main area of expertise is Natural Interaction, which is concerned with providing natural and tangible interfaces to multimedia information systems. In the context of



(a) (b)

Fig. 1. (a) Intelligent visual information systems, via a combination of fixed and wearable sensors, will analyze and reason about the interactions of visitors in the physical space. (b) At interactive tabletops placed throughout the museum environment, visitors will be presented with personalized, interactive suggestions about other possible exhibits.

MNEMOSYNE, figure 1b illustrates a scene in which museum visitors are provided with a customizable, personalized interaction surface with which they can interact with museum assets.

The other area of technical expertise key to the MNEMOSYNE project is Computer Vision. In figure 1a is shown a broader view of a museum interior. Using a combination of fixed and worn sensors, intelligent visual information systems will be built to interpret and profile visitors to the museum. In addition to providing a secure environment for physical museum assets, these visual systems will provide the information necessary for profiling the user and customizing the interactive terminals placed throughout the museum environment.

The MNEMOSYNE project is in its initial planning phase in which we have begun consolidating our existing work in related fields. In this article we describe some of the decisions we have made and the direction we will take MNEMOSYNE over the next three years. In the next section we discuss the state-of-the-art in interactive multimedia museums. In section 3 we discuss issues related to personalization in multimedia museum displays. Section 4 contains a discussion of the computer vision techniques we will bring to bear on the problem of visual profiling of museum guests. We conclude with a discussion in section 5 and indications of the trajectory the MNEMOSYNE will follow in the years to come.

2 Multimedia Interactive Museums

The presence of videos, interactive applications and detailed websites not only help visitors to manage the complexity of exhibited content, but also improves the persistence in memory of concepts through the use of different senses. While visitors are becoming acquainted with the presence of computers displaying various media in an exhibition or museum, these devices are usually placed in hidden

Fig. 2. Interactive surfaces engage museum visitors in ways traditional museum exhibits cannot

corners and rarely provide an interface designed for maximizing knowledge transfer and taking account of user experience: when it comes to multimedia, the offer is often limited to variants of common web sites.

Our perspective is centered on user interaction with digital devices specifically designed to reduce cognitive effort: this means that the interface design allows users to interact and to actually concentrate on content rather than thinking about how to use the interface. Focusing on content means that a wider point of view can be conveyed through the use of senses rarely used in a common PC environment, such as a proactive combination of touch, hearing and sight. Having this paradigm in mind, the cognitive load usually carried by the interface can be shifted to data, thus raising the level of complexity of transmitted information and achieving the goal of a technical and scientific communication of exhibit details.

The User Interface (UI) is the main contact point between users and computers: it is what users see, hear and touch; it is the perceptive representation of the communication channel which users use to communicate to the system and vice-versa. The research of new kinds of Natural Human-Computer Interaction systems is a growing field in computer science which aims to develop more intuitive, efficient, easy-to-use and satisfactory interfaces [6]. At the same time, researchers and companies (Microsoft with Surface, Nintendo, etc.) are trying to design and develop new devices to aid this process. In [22], Turk and Robertson divide User Interfaces into 3 different categories:

- **Perceptive user interfaces** which provide computers with human-like perceptual capabilities, so that implicit and explicit information about users and their environment can be conveniently acquired. The machine is able to see, hear or sense;
- **Multimodal user interfaces** which exploit multiple forms of input and/or output. In multimodal UI, various modalities can be used independently or simultaneously; and
- **Multimedia user interfaces** which are focused on media, such as text, graphics, video and/or sound.

Natural Human-Computer Interaction (NHCI) can be seen as a fusion of these kind of interfaces: its task is to make user interfaces more natural by taking into account the ways in which people naturally interact with each other and with the world. Among such systems, interactive surfaces that allow multiple users' touch are suitable for collaborative applications, especially in educational, professional and entertainment environments [13].

In MNEMOSYNE we have chosen the Natural Interaction paradigm because it allows us to design multimedia displays that require no user familiarity or training in order to use. They can begin interacting with their personalized, interactive multimedia presentation by simply walking up to one of the MNEMOSYNE multitouch tables present in the museum exhibit. The challenge will be in designing interfaces, multimedia content and user profiling primitives that interoperate in a seamless and natural way.

3 Personalizing the Multimedia Museum Experience

In recent years, the purpose of museums has shifted from merely providing static information to delivery of personalized interactive contents: before, it was very difficult for museum visitors to find the right information at the right time and at the right level of detail. The idea is then to turn the museum monologue into a user-centred dialog between the museum and its visitors [8]. This solution is what we call personalized multimedia tours. What distinguishes these from the traditional "static" ones is the exploitation of a user-model that represents the characteristics of each user [9]. The information stored in the user-model is exploited in the process of content generation to describe or recommend objects potentially relevant to users.

Personalized applications have exploited recent research results in recommender systems, information retrieval and data mining which provide important solutions for a user-centered interactive information exchange between museums and their visitors. The user information can be inferred implicitly by observing visitor behaviour or during their interactions with the multimedia device; it can also be provided explicitly by the users [8]. The main challenge is to analyse interests and preferences without demanding them to express them explicitly: it is more desirable to start offering recommendations to visitors as soon as possible, hence minimizing intrusiveness to the users [17]. These types of solutions are quite complex and have been developed in the context of academic research. For example, the Wearable Computer (MIT Media Lab) provides audio and visual narration adapting to the users interest from him/her physical path in the museum and length of stops [20]. The PEACH project [18] developed a PDA-based museum tour application, whose content is adapted to the visitor, is location-aware and only available in certain locations in the museum. In CHIP [24] there is an automatic user-model that collects user interests from his or her interactions. Content-based recommendations are then exploited to recommend both artwork and art concepts that might be of interest.

Personalization methods are often classified into main categories such as collaborative filtering, content-based filtering, and hybrid approaches [1]. Collaborative methods divide visitors into groups that have similar known preferences and then recommend to a new visitor those items that were most liked by the group to which he or she belongs. The problem is that new works in the collection will not be recommended until a substantial number of users have rated them.

Content-based methods analyze the common features among the items a visitor liked and recommends those items that have similar features. The main problem of applying content-based techniques for museum tour personalization is that both automatic feature extraction from graphical images and manual assignment of these features are difficult so that a recommender system usually has a rather limited set of features, which may not tell about some qualities of some artwork. An additional problem here is that two different pieces with the same set of features (with similar values), can not be distinguished by the recommender system.

For MNEMOSYNE, we have chosen a Hybrid approach that combines collaborative and content-based methods: we combine recommendations produced separately by content-based and collaborative techniques to develop a generic model that includes elements of both types of techniques.

4 Active Vision for Visitor Profiling

Several application domains, including museum security and surveillance, rely on large number of video sequences captured using a combination of cameras of various types: fixed-lens, steerable pan-tilt-zoom, thermals, omnidirectional, and handheld sensors [11]. Recent progress in camera hardware and communication technologies has led to an evolution of camera networks into networks of smart cameras and highly capable mobile imaging systems. These open up novel opportunities for scientific discovery in image analysis, especially in wide area scene analysis. Wide area scene analysis builds upon multi-view image analysis, which is an active sub-area of computer vision which use visual data captured via camera networks and has its own unique challenges, including:

- The ability to integrate information over a wide area;
- Cooperation between camera;
- Active control of the network; and
- Scalability.

These advances in recent years have created a unique opportunity for the MNEMOSYNE project: to exploit state-of-the-art computer vision techniques in heterogeneous sensor networks as well as the installed video surveillance infrastructure in modern museum environments in order to perform remote visitor profiling in terms of what museum pieces they exhibit interest in, how long they linger before a particular display, and what course they follow through the museum environment. In MNEMOSYNE we will leverage our existing work in

computer vision and video surveillance to perform this "visual profiling" of museum guests. In this section we describe some of the architectural aspects of the camera network and systems that we will use to implement museum visitor profiling in the MNEMOSYNE project.

4.1 Building Blocks

In order to profile visitors, the system will use a network of fixed (master) cameras, and a number of active PTZ (slave) cameras. We have described in previous work several components to:

- **Accurately detect and track visitors [3]**
  This module will be responsible for keeping track of visitors and their route as then move through the museum. Our sensor network will incorporate a combination of fixed and active sensors, which complicates this aspect of the system.
- **Capture high-quality face imagery [12]**
  This component will allow us to associate an "identity" in the form of a face with each tracked visitor.
- **Recognize actions of visitors [4]**
  Simple actions like pointing or grouping, when accurately detected, are reasonable indicators of visitor interest.

The first component is one of the most investigated in the literature on computer vision and video surveillance. In the heterogeneous sensor case, however, there is significantly less work to build on. In this case two cameras are typically associated in a master-slave relationship: the master camera is kept stationary and set to have a global view of the scene so as to permit it to track several entities simultaneously. The slave camera is used to follow the target trajectory and generate close-up imagery of the entities driven by the transformed trajectory coordinates, moving from target to target and zooming in and out as necessary. The solutions proposed in [2] [25] do not require direct calibration but impose some restrictions in the setup of the cameras. The viewpoints between the master and slave camera are assumed to be nearly identical so as to ease feature matching. In [25], a linear mapping is used that is computed from a look-up table of manually established pan and tilt correspondences. In [2], a look-up table is employed that also takes into account camera zooming. In [19], it is proposed a method to link the foot position of a moving person in the master camera sequence with the same position in the slave camera view. The methods proposed by [14], [15] require instead direct camera calibration, with a moving person and calibration marks.

4.2 Preliminary Results: Scene Learning and Visitor Tracking

In this preliminary phase of the project the focus is on establishing at frame-rate the time variant mapping between PTZ cameras present in a network as they redirect the gaze and zoom to acquire high resolution images of moving targets

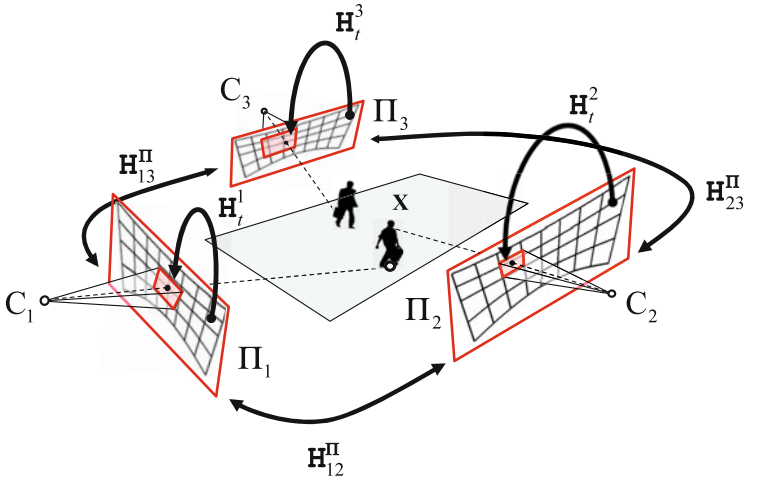


Fig. 3. Pairwise relationships between PTZ cameras for a sample network of three cameras

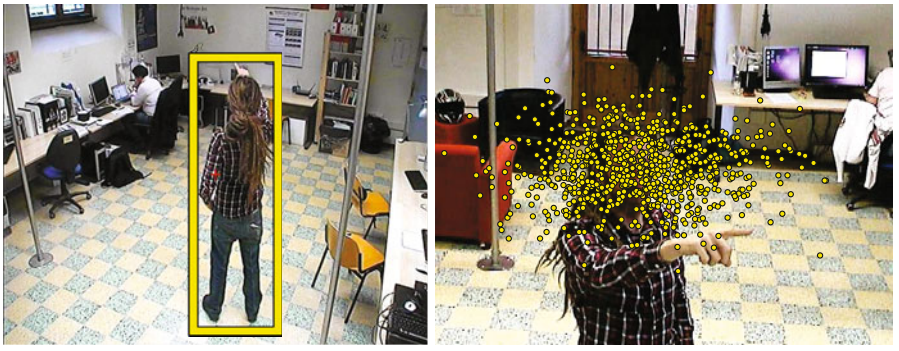
for profiling purposes. This is critical in the MNEMOSYNE scenario because any system must be easily re-deployable in new environments, and should also work in semi-stable environments in general (exhibits change).

We build upon the work [7] which exploits a prebuilt map of visual 2D landmarks of the wide area to support multi-view image matching. The landmarks are extracted from a finite number of images taken from a non calibrated PTZ camera. At run-time, landmarks that are detected in the current PTZ camera view are matched to those of the base set in the map. The matches are used to localize the camera with respect to the scene and hence estimate the position of the target body parts.

According to [7], cameras with an overlapping field of view can be set in a master-slave relationship pairwise. According to this, given a network of M PTZ cameras $\mathbf{C}_i$ viewing a planar scene, $\mathcal{N} = \{\mathbf{C}_i^s\}_{i=1}^M$, at any given time instant each camera can be in one of two states $s \in \{\text{MASTER}, \text{SLAVE}\}$.

As shown in Fig. 3 the three reference planes Π_1, Π_2, Π_3 observed respectively by the cameras $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$ are related to each other through the three homographies $\mathtt{H}_{12}^{\Pi}$, $\mathtt{H}_{13}^{\Pi}$, $\mathtt{H}_{23}^{\Pi}$. Instead, at time t the current image plane is related to the reference plane through the homographies $\mathtt{H}_t^1$, $\mathtt{H}_t^2$ and $\mathtt{H}_t^3$. If the target $\mathbf{X}$ is tracked by $\mathbf{C}_1$ (acting as MASTER) and followed in high resolution by $\mathbf{C}_2$ (acting as zooming SLAVE), the imaged coordinates of the target are first transferred from Π_1 to Π_2 through $\mathtt{H}_{12}^{\Pi}$ and hence from Π_2 to the current zoomed view of $\mathbf{C}_2$ through $\mathtt{H}_t^2$. Referring to the general case of M distinct cameras, once $\mathtt{H}_t^k$ and $\mathtt{H}_{kl}^{\Pi}$, $k \in 1..M$, $l \in 1..M$ with $l \neq k$ are known, the imaged location of a moving target tracked by a master camera $\mathbf{C}_k$ can be transferred to the zoomed view of a slave camera $\mathbf{C}_l$ according to:

$$\mathtt{T}_t^{kl} = \mathtt{H}_t^l \cdot \mathtt{H}_{kl}^{\Pi} \tag{1}$$



(a) (b)

Fig. 4. Example of a frame analyzed with the proposed technique. (a): Master camera view: the target is detected by background subtraction. (b): Slave camera view: the particles show the uncertainty of the head position of the target.

Under the assumption of vertical stick–like targets moving on a planar scene the target head can be estimated directly by a planar homology [23,10] as shown in Fig. 4.

In our preliminary experiments we have seen that in indoor environments we are able to quickly learn a map of visual landmarks that allow our cameras to orient themselves in the environment. These maps can then be used to accurately locate and track humans in the 3D environment.

4.3 Preliminary Results: Head Localization

In order to extract more information more meaningful to the task of visitor profiling, we have also concentrated on extracting head positions from the views of the slave-camera. In MNEMOSYNE this will be used for focusing higher-level analysis modules on heads in order to estimate, for example, *where* the visitor is looking.

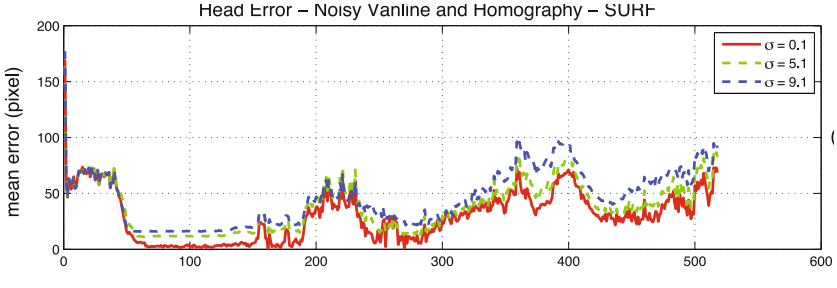


Fig. 5. Head localization accuracy for the test sequence shown in Fig. 4

To evaluate the head localization error in the slave camera we corrupt the two pairs of parallel lines needed to estimate the vanishing line and the four points needed to estimate the homography $\mathtt{H}_{12}^{\Pi}$, with a white, zero mean, Gaussian noise with standard deviation between 0.1 and 9 pixels. This procedure was repeated 1000 times and averaged over trials. Plots of the mean error in head localization are reported in Fig. 5. As it can be seen, after a brief transient (necessary to estimate the initial camera pose), the mean error falls to small values and grows almost linearly as the focal length increases.

5 Discussion

MNEMOSYNE is a three-year project funded by the Region of Tuscany and the European Commission whose main goal is to find new solutions for adaptive user-profiling in interactive museum contexts. The unique and novel aspect of the MNEMOSYNE project is that it exists precisely at the point where interactive museums and computer vision intersect. The tools of computer vision will give us complete coverage of each visitor's pattern of visitation. With this dense flow of information about each visitor, we will be able to build a more accurate profile and customize the experience for them on-the-fly. This type of profiling also opens up the door to suggest other exhibits, other museums, or to create networks of MNEMOSYNE exhibits which are interlinked to create a web of interactive, multimedia museums.

Preliminary experiments with existing computer vision systems have only just begun in a laboratory setting. Ongoing work is concentrating primarily on (i) generalizing the visual landmark mapping system to work in indoor, crowded environments; and (ii) incorporating mobile, worn and/or carried sensors (e.g. smartphones) into our sensor network model, fusing information with these streams in order to accurately associate visitor observations without the need for accurate tracking.

Acknowledgment. This work is supported by a grant from *La Regione Toscana* and the European Commission.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17 (2005)
2. Badri, J., Tilmant, C., Lavest, J.-M., Pham, Q.-C., Sayd, P.: Camera-to-Camera Mapping for Hybrid Pan-Tilt-Zoom Sensors Calibration. In: Ersbøll, B.K., Pedersen, K.S. (eds.) SCIA 2007. LNCS, vol. 4522, pp. 132–141. Springer, Heidelberg (2007)
3. Bagdanov, A.D., Dini, F., Del Bimbo, A., Nunziati, W.: Improving the robustness of particle filter-based visual trackers using online parameter adaptation. In: Proc. of IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS). IEEE Computer Society, London (2007), http://www.micc.unifi.it/publications/2007/BDDN07a

4. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Effective codebooks for human action categorization. In: Proc. of ICCV Int'l Workshop on Video-oriented Object and Event Classification (VOEC), Kyoto, Japan (September 2009), http://www.micc.unifi.it/publications/2009/BBDSS09a
5. Baraldi, S., Bimbo, A.D., Landucci, L.: Natural interaction on the tabletop. Multimedia Tools and Applications (2008)
6. Baraldi, S., Bimbo, A.D., Landucci, L., Torpei, N.: Entry: Natural interaction. In: Szsu, M.T., Liu, L. (eds.) Encyclopedia of Database Systems. Springer, Heidelberg (2008)
7. Bimbo, A.D., Dini, F., Lisanti, G., Pernici, F.: Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks. Computer Vision and Image Understanding 114(6), 611–623 (2010), special Issue on Multi-Camera and Multi-Modal Sensor Fusion
8. Bowen, J., Filippini-Fantoni, S.: Personalization and the web from a museum perspective. In: Proceedings of the 2004 Museums and the Web Conference (2004)
9. Brusilovsky, P., Maybury, M.T.: From adaptive hypermedia to the adaptive web. Communications of the ACM 45(5) (2002)
10. Criminisi, A., Reid, I., Zisserman, A.: Single view metrology. International Journal of Computer Vision 40(2), 123–148 (2000)
11. Del Bimbo, A., Dini, F., Grifoni, A., Pernici, F.: Pan-tilt-zoom camera networks. In: Aghajan, H., Cavallaro, A. (eds.) Multi-Camera Networks: Principles and Applications, Academic Press (2009)
12. Del Bimbo, A., Dini, F., Lisanti, G.: A real time solution for face logging. In: Proc. of International Conference on Imaging for Crime Detection and Prevention, ICDP (2009), http://www.micc.unifi.it/publications/2009/DDL09
13. Fikkert, F.W., Hakvoort, M., van der Vet, P.E., Nijholt, A.: Experiences with interactive multi-touch tables. In: Proceedings of INTETAIN (2009)
14. Horaud, R., Knossow, D., Michaelis, M.: Camera cooperation for achieving visual attention. Mach. Vis. Appl. 16(6), 1–2 (2006)
15. Jain, A., Kopell, D., Kakligian, K., Wang, Y.F.: Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention. In: CVPR 2006: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 537–544. IEEE Computer Society, Washington, DC, USA (2006)
16. Norman, D.: The Invisible Computer. The MIT Press (1998)
17. Pechenizkiy, M., Calders, T.: A framework for guiding the museum tour personalization. In: Proceedings of PEACH 2007 (2007)
18. Rocchi, C., Stock, O., Zancanaro, M., Kruppa, M., Kruger, A.: The museum visit: Generating seamless personalized presentations on multiple devices. In: Proceedings of the 2004 Conference on Intelligent User Interfaces (2004)
19. Senior, A., Hampapur, A., Lu, M.: Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration. In: IEEE Workshop on Applications on Computer Vision (2005)
20. Sparacino, F.: The museum wearable: real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences. In: Proceedings of Museums and the Web, MW 2002 (2002)
21. Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R., Pentland, A.: Augmented reality through wearable computing. Presence 6(4) (1997)
22. Turk, M., Robertson, G.: Perceptual user interfaces. Communications of the ACM (2000)

23. Van Gool, L., Proesmans, M., Zisserman, A.: Grouping and invariants using planar homologies. In: Workshop on Geometrical Modeling and Invariants for Computer Vision. Xidian University Press (1995)
24. Wang, Y., Aroyo, L., Stash, N., Sambeek, R., Schuurmans, Y., Schreiber, G., Gorgels, P.: Cultivating personalized museum tours online and on-site. Interdisciplinary Science Reviews 34(2) (2009)
25. Zhou, X., Collins, R., Kanade, T., Metes, P.: A master-slave system to acquire biometric imagery of humans at a distance. In: ACM SIGMM 2003 Workshop on Video Surveillance, pp. 113–120 (2003)