VMASS: Massive Dataset of Multi-camera Video for Learning, Classification and Recognition of Human Actions

Marek Kulbacki, Jakub Segen, Kamil Wereszczyński, and Adam Gudyś

Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008 Warsaw, Poland
{mk,js,kw,agudys}@pjwstk.edu.pl

Abstract. Expansion of capabilities of intelligent surveillance systems and research in human motion analysis requires massive amounts of video data for training of learning methods and classifiers and for testing the solutions under realistic conditions. While there are many publicly available video sequences which are meant for training and testing, the existing video datasets are not adequate for real world problems, due to low realism of scenes and acted out human behaviors, relatively small sizes of datasets, low resolution and sometimes low quality of video.

This article presents **VMASS**, a dataset of large volume high definition video sequences, which is continuously updated by data acquisition from multiple cameras monitoring urban areas of high activity. The VMASS dataset is described along with the acquisition and continuous updating processes and compared to other available video datasets of similar purpose. Also described is the sequence annotation process. The amount of video data collected so far exceeds 4000 hours, 540 million frames and 2 million recorded events, with 3500 events annotated manually using about 150 event types.

1 Introduction

The intelligent event recognition is one of the most important issues of computer vision [2–7]. To construct Intelligent Video Analytics systems massive video datasets with annotations are necessary, as training and test data for machine learning processes. The following criteria can help to assess usefulness of a video dataset towards a particular task: realism, quality described by multiple parameters, variety of stored events and actions, actors, scenes and external conditions, and the annotation method.

Realism of dataset is a property of recording not pre-arranged events in a not pre-arranged scene, where subjects are incidental and not conscious of being recorded. This property relates to the degree of how well the recorded data represents the real world. System trained on non-realistic dataset usually fail in the real world. Few existing datasets satisfy this criterion, since most rely on a small number of actors performing pre-arranged actions [1, 4] or use movie scenes [2]. The datasets that possess this property are [5], CAVIAR [8] and VIRAT [3].

N.T. Nguyen et al. (Eds.): ACIIDS 2014, Part II, LNAI 8398, pp. 565–574, 2014.

Quality of a video dataset can be represented numerically by parameters: image resolution, image or video compression level and the number of frames per second (FPS). Within the bounds of present technology, a good quality video dataset will have Full HD resolution (1920x1080), compression corresponding to JPEG 10-20% level without motion compensation and at least 18 FPS. An example of a good quality video dataset is VIRAT [3].

Variety of conditions relates to how broad are the ranges of conditions or parameters within which the system operates or represented objects. It includes the quantity of events or actions, the number of actors and scenes (in background object appearance and activity aspects), the external conditions such as as season, time of day, weather, illumination or stability of the background. Lack of variety in existing datasets is usually represented by short recording time, small number of pre-arranged scenes and actions performed by few actors, or clear weather and good illumination.

Annotation method effects the usefulness of the dataset as training and test data. Multiple types of annotations are needed: (1) Annotation of time segments; (2) Annotation of objects as used by by KTH [1], Weitzmann [4] or TRECVID [9]; (3) Multiple annotations of objects as in Virat [3]; (4) Hierarchical annotations, for example an action divided into smaller structures (the authors don't know of its use before **VMASS**).

Calibration of a view in a video dataset, or camera calibration, concerns the information that makes possible to relate geometrically different views, different cameras and assess true object sizes. While calibration may not be necessary for a dataset with a single view, such as most of the existing datasets, it becomes critical with video segments from different cameras or multiple views. Obtaining the calibration information under minimal operator's involvement is not trivial, but even more difficult is a continuous computing of the calibration parameters with cameras in motion. **VMASS** contains calibration parameters for every view, that are computed in real time during the acquisition.

Size of a video dataset can be expressed by several parameters such as: (1) size of data (GB); (2) length of the recording; (3) number of frames. Existing available datasets range from several to under 100 hours, while **VMASS** contains more than 4.000h of video data.

2 VMASS Dataset

The Polish-Japanese Institute of Information Technology (PJIIT) in its Bytom branch conducts research in the areas of human motion analysis and video based activity recognition. As most projects are based on machine learning techniques,

large amounts of video data are needed for the training and testing. The VMASS system and the acquired massive video dataset were build to fulfill this need. The following subsections describe properties of the VMASS system within the guidelines of the criteria introduced in Section 1.

2.1 Realism

The cameras used for vision data acquisition were located in the center of Bytom, Poland. The acquisition system collects video data, outdoor, recording real situations without any control of the scene or events. This classifies it as a high Realism system according to the criteria described in Section 1.

Fig. 1. A typical VMASS video frame as an example for data realism

2.2 Quality

The four video cameras used for the acquisition of video dataset are High Definition IP auto-dome cameras made by Axis Communications, model Q6035E. Technical details of this equipment are:

- Image sensor resolution: p1080 (1920 x 1080)
- Zoom scale: 1x - 20x
- Pan / Tilt position setting precision: about 0.01 deg.
- Frame rate in various resolutions: 1920x1080 MJPEG: 25fps; 1024 x 768: 50 fps; 640 x 480: 50 fps.

An important part of acquisition process is the efficient delivery of video data to a storage center. This task is being realized by optical and 6^{th} category level

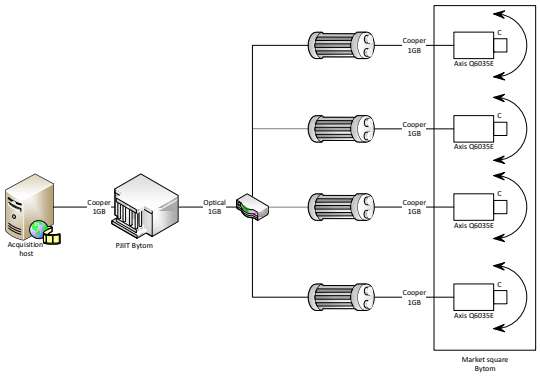

Fig. 2. VMASS video acquisition network

copper network connecting all cameras with CVL with guaranteed bandwidth on level 1GBps.

The quality of the acquired data depends on image resolution, frame-rate and inversely on the compression level. The video parameters of the recorded sequences are: resolution 1920 x 1080, frame-rate 18-27 fps, JPEG compression on 10-13%, video compression: no compression of motion areas (see detailed description in next section). Six to eight hours of video per day are recorded from each camera.

2.3 Variety of Conditions

Variety of conditions in the VMASS system is illustrated by examples in Figure 3.

Fig. 3. Variety of conditions in the VMASS

The range of conditions used in the VMASS dataset is shown in Table 1.

Table 1. Variety of conditions in the VMASS dataset

Condition type	**Range**	**Examples**
event	1 – 150	walk, talking, bicycle, run, run away, photograph, play, sit, downfall, get out of building, carry luggage, waiting, smoke, phone,...
actor	1 – 40	man, woman, pair, threesome, group, crowd, dog, luggage, car, truck, excavator, ...
scene	1 – 52	market square in bird-eye like view, church entrance, supermarket entrance, fountain, open-air cafe, road crossing, trees, street, building entrance, ...
External conditions		
seasons	1 – 4	spring, summer, fall, winter
day times	daylight	morning, noon, evening
weather	any	sunny, windy snowy, rainy, cloudy (many level of cloud cover),...
illumination	visible	depending on weather conditions, daytime, season and so on.
background stability	0 - moderate motion	stable,camera in motion, appearing and disappearing objects in background, birds, leafs, raindrops, etc existing in background.

2.4 Annotation Method

The annotation method of VMASS can be described as Hierarchical, Parallel, Multidimensional and Flexible (HPMF).

Hierarchical means that each annotation can have sub-annotations and each of them sub-sub-annotations and so on (e.g. event "meeting" can have sub-annotations: "greetings", "chat", "walk away" and "greetings" could have sub-annotations: "waving", "handshake", "kiss"). This annotation tree could be as deep as needed.

Parallel means that on the same time and the same object different set of annotations can be defined. (e.g. given person could attend an event "meeting" with person A and event "jostle" with person B; in the same time other two persons can attend event "play"). It means also, that one event could contain more than one tag of each dimension.

Multidimensional means that one event/action could be annotated in several perspectives (called dimensions) with different tags. For each dimension there is defined other tag set. Such annotations are 4-dimensional (action, behavior, actor and scene).

Flexible means that depth of annotation hierarchy is unlimited, dimension count and tags dictionary for each dimension can be modified.

A software system for adding the annotations to video sequences called Annotation Editor has been developed at PJIIT Bytom. Figure 4 shows a snapshot of the Annotation Editor window.

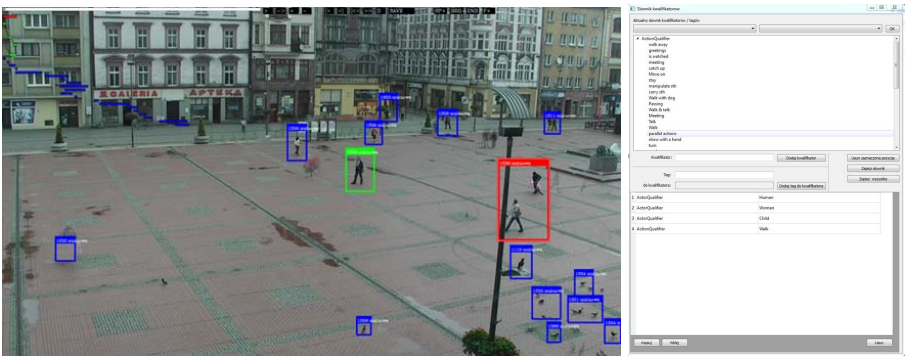

Fig. 4. Screenshots from acquisition application

On the left side of the screen starting from the top:time-line (white); Annotations (blue bars mean not selected and red bar means selected); Sub-annotation of selected annotation - green bars below selected activity.

Rectangular windows: Red window - Selected motion areas (could more then one); Blue window - motion areas not connected with selected activity; Green window - motion areas connected with selected activity.

2.5 Comparison to Other Datasets

Results of comparison between VMASS and other published datasets are summarized in Tables 2, 3, 4 and 5.

Table 2. Realism

	KTH	Wiezmann	HOHA 1	TRECVID	VIRAT	VMASS
incidental events, scenes, actors	No	No	Partial	Yes	Yes	Yes
realistic environment	No	No	No/Partial	Partial	Partial	Yes

Table 3. Quality

	KTH	Wiezmann	HOHA 1	TRECVID	VIRAT	VMASS
resolution (wxh)	160x120	180x144	540x240	720x576	1920x1080	1920x1080
frame rate	N/A	N/A	N/A	N/A	N/A	18-27 fps
frame compression level	N/A	N/A	N/A	N/A	N/A	10-13% (JPEG)
video compression	N/A	N/A	N/A	N/A	N/A	lossless MJPEG
zoom	fixed	fixed	fixed	fixed	N/A	1-20 (optical)
actors height in pixels	80-100	60-70	100-120	20-200	20-180	20-1000

Table 4. Variety of conditions

	KTH	Wiezmann	HOHA 1	TRECVID	VIRAT	VMASS
count of events and action types	6	10	8	10	23	150(e), 400(a)
avg. count of samples per class	100	9	85	3-1670	10-1500	5-1200
count of scene types	N/A	N/A	many	5	17	over 50
seasons	N/A	N/A	N/A	N/A	N/A	4
weather	N/A	N/A	N/A	N/A	N/A	many
illumination	N/A	N/A	N/A	N/A	N/A	many
background stability	Yes	Yes	cam. motion	Yes	cam. motion	varying[1]

Table 5. Annotation system and data size

	KTH	Wiezmann	HOHA 1	TRECVID	VIRAT	VMASS
annotation type	1 per obj	1 per obj	1 per obj	1 per obj	multiple	HPMF
dataset size [h]	few seq.	few seq.	few seq.	few seq.	28	4.000
dataset size [million frames]	N/A	N/A	N/A	N/A	2	540
approximate recorded events count [thousands]	0.6	0.09	0.68	10	23	3.000

[1] Stable, camera motion, appearing and disappearing objects in background, birds, leafs, raindrops, etc existing in background.

3 Overview of Implementation of the VMASS System

The main parts of the data collection process are video acquisition, compression, camera view registration, storage and annotation. The first three parts of the process are automatic, the fourth part - annotation is manual. A video stream acquired from a camera is a sequence of JPEG compressed images. Each frame is additionally compressed using a method of object selective compression, and stored in a compressed form. Each distinct camera view is registered with a map with a reference ground plane. To add annotations to the sequence, the stored video scenes are retrieved and processed manually using the Annotation Editor. The object selective compression method and the view registration are described in the following subsections.

3.1 Object Selective Compression

Object selective compression allows the system to store moving objects at higher accuracy and non-changing background at a lower accuracy. For each frame the regions containing moving objects are stored as new information while the background representation is taken from the previous frame, disregarding minor background changes. The moving objects are preserved at their acquisition quality while the background is stored less accurately since minor changes in the background are omitted, which results in a high degree of compression. The process of object selective compression consist of: (1) Identifying background and storing it as B; (2) Detecting motion rectangles for each frame; they will be called *crumbles* and marked with a letter C; (3) Storing the sequence as: $BCCCCCBCCCB$; (4) Rebuilding a frame on demand as $B_l + C_i$ where i is the index of demanded frame and B_l is last B before C_i.

This process uses the background subtraction method based on Adaptive Gaussian Mixture Model proposed by P. KaewTraKulPong and R. Bowden [12] and Z. Zivkovic [13]. These models focus on one pixel color change between consecutive frames and basing on this change try to estimate most probable background color of this point using an idea that this color is the one which stay

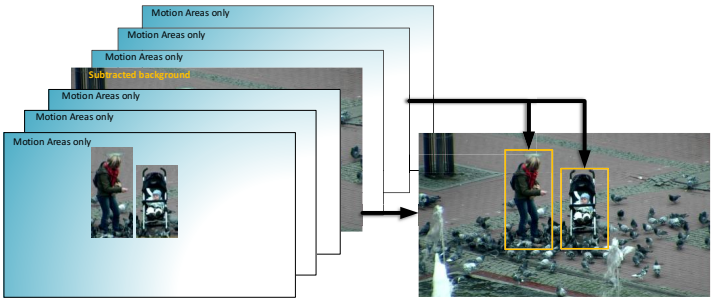

Fig. 5. Object selective compression

longer and more static on given pixel. Once background is known, silhouettes are retrieved, basing on observation differences between background and given frame. For eliminated fragmentation of retrieved object, mixture of filtering and thresholding methods are used. When silhouettes are retrieved for each frame, motion areas could be discovered, basing on calculating motion history, described by Bradsky and Davis in 2002 in [14].

3.2 Camera View Registration

The view registration, which is the mapping of the image elements that lie on the ground plane to positions on the reference frame, is done from known camera calibration parameters. A camera calibration method has been developed as a part of PJIIT-MIS project. It updates the calibration parameters when the camera pan and tilt angles change. The camera calibration is specified by two matrices: intrinsic and extrinsic.

Intrinsic Camera Parameters. Camera intrinsic parameters are: focal length f, skew s, coordinates of principal point $P = (p_u, p_v)$, scale factors in horizontal and vertical directions k_u,k_v. The intrinsic parameters are computed using the Hartley's algorithm [10].

Extrinsic Camera Parameters. The extrinsic parameters are given by the camera pose, which consists of its orientation angles and the position. Pose is represented by translation vector $T_{3\times1}$ and rotation matrix $R_{3\times3}$. Pose is obtained from the correspondence between coordinates of a scene object and its image. The algorithm used by the calibration module is a version of the POSIT method for coplanar points (Oberkampf et. Al, 1996) [11].

Maintaining the Calibration. The initial camera pose estimation (or its extrinsic calibration) is calculated from a set of image points and their corresponding points on the reference plane. Such initial calibration is done for a small set of camera poses. The calibration parameters for the other poses are computed recursively starting from this initial set. For a rotating camera image stitching is used to obtain a sequence of camera poses, following Hartley's algorithm [10]. For each pair of successive images a set of common points is found that allows computing the pose of the second image from the pose of the first image.

4 Conclusion

A new approach and a system VMASS for constructing massive video datasets with annotations has been presented. Such datasets are needed for training and testing methods of classification and recognition of human activities. VMASS dataset represents a significant improvement comparing to the existing datasets with a similar purpose, due to its realism, image quality, variety of actions and

situations, hierarchical annotation protocol and a very high volume of stored data. The **VMASS** dataset has become a essential resource for computer vision projects at PJIIT. We expect it will be useful to others and we plan to make it publicly available.

Acknowledgments. This work was supported by projects NN 516475740 from the Polish National Science Centre and PBS I ID 178438 path A from the Polish National Centre for Research and Development.

References

1. Schuldt, C., Laptevand, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR (2004)
2. Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV, pp. 1–8 (2007)
3. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J.T., Mukherjee, S., Aggarwal, J.K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011, pp. 3153–3160 (2011)
4. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. PAMI 29(12), 2247–2253 (2007)
5. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the Wild". In: CVPR 2009 (2009)
6. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. CVIU 104(2), 249–257 (2006)
7. Ke, Y., Sukthankar, R., Hebert, M.: Volumetric Features for Video Event Detection. IJCV 88(1) (2010)
8. Fisher, R.B.: The PETS04 Surveillance Ground-Truth Data Sets (2004)
9. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: MIR 2006 (2006)
10. Hartley, R.I.: Self-Calibration from Multiple Views with a Rotating Camera. In: Eklundh, J.-O. (ed.) ECCV 1994, Part I. LNCS, vol. 800, pp. 471–478. Springer, Heidelberg (1994)
11. Oberkampf, D., DeMenthon, D.F., Davis, L.S.: Iterative Pose Estimation Using Coplanar Feature Points. Computer Vision and Image Understanding 63(3), 495–511 (1996)
12. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Proc. 2nd European Workshop on Advanced Video-Based Surveillance Systems (2001)
13. Zivkovic, Z.: Improved Adaptive Gaussian Mixture Model for Background Subtraction. In: International Conference Pattern Recognition, UK (August 2004)
14. Davis, J.W., Bradski, G.R.: Motion Segmentation and Pose Recognition with Motion History Gradients. Machine Vision and Applications (2002)