Unsupervised Surveillance Video Retrieval based on Human Action and Appearance

David Gerónimo and Hedvig Kjellström
Computer Vision and Active Perception Lab
KTH Royal Institute of Technology
Stockholm, Sweden
{dgero,hedvig}@kth.se

***Abstract*—Forensic video analysis is the offline analysis of video aimed at understanding what happened in a scene in the past. Two of its key tasks are the recognition of specific actions, e.g., walking or fighting, and the search for specific persons, also referred to as re-identification. Although these tasks have traditionally been performed manually in forensic investigations, the current growing number of cameras and recorded video leads to the need for automated analysis. In this paper we propose an unsupervised retrieval system for surveillance videos based on human action and appearance. Given a query window, the system retrieves people performing the same action as the one in the query, the same person performing any action, or the same person performing the same action. We use an adaptive search algorithm that focuses the analysis on relevant frames based on the inter-frame difference of foreground masks. Then, for each analyzed frame, a pedestrian detector is used to extract windows containing each pedestrian in the scene. For each detection, we use optical flow features to represent its action and color features to represent its appearance. These extracted features are used to compute the probability that the detection matches the query according to the specified criterion. The algorithm is fully unsupervised, i.e., no training or constraints on the appearance, actions or number of actions that will appear in the test video are made. The proposed algorithm is tested on a surveillance video with different people performing different actions, providing satisfactory retrieval performance.**

I. Introduction

In the last decades, communications and storage technologies have undergone a huge progress, experiencing a rapid increase in the amount and quality of data generated, acquired and stored. This is certainly the case in the field of visual surveillance. Nowadays, the amount of cameras aimed at investigating crimes or causes of accidents has raised significantly, establishing themselves as a relevant tool in the task of solving crimes such as the recent Boston bombings. The field that analyzes such videos in the search for specific persons, recognize actions and interactions is called forensic video analysis [1]. Traditionally, such analysis has been carried out manually. However, during the last years, extensive research has been developed in content-based image and video mining in order to automatize the process. Related topics such as person detection [2] and action classification [3] in videos have progressively gained attention from the researchers. In spite of this, video forensics is still largely carried out manually, not only for short videos but also for videos of thousands of hours. Although this processing is nowadays still feasible with huge human effort, the growing number of cameras and recorded hours will make the video forensics task much harder in the future, potentially leaving events or objects undiscovered.

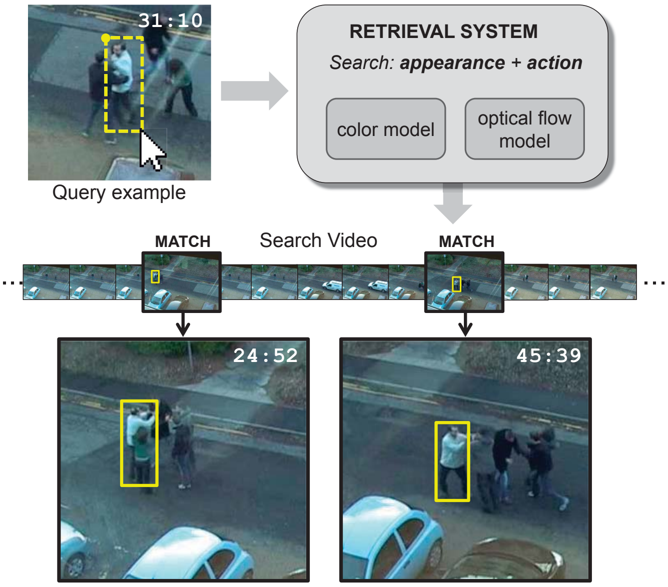

Fig. 1. Proposed surveillance image retrieval system. Given a query bounding box, the system outputs matches to the query according to its appearance and/or action in a search video.

Video surveillance systems perform two key tasks, recognizing actions and searching for specific persons (e.g., re-identification), among others such as people tracking or anomaly detection. In the context of forensics we refer to these two tasks as action retrieval and appearance retrieval. Action retrieval is focused on searching for clips in which a person performs a given action, e.g., running or fighting, as defined by a textual or visual query. Appearance retrieval is focused on the same kind of search but constrained to appearance, e.g., similar clothing or biometric measures.

Although forensic video analysis in surveillance can also deal with the same scenarios as traditional online scene monitoring, e.g., shopping malls or underground stations, the nature of the analysis is different: the former is focused on offline processing larger amounts of data, usually in the context of a crime investigation. Such a system will work as follows. First, an operator selects a specific person either manually or automatically attending to some evidence. Then, the system will automatically retrieve instances of such a person in the videos. The benefits of such an automated retrieval are many: it reduces the amount of data to investigate the places and times in which the person appears or performs the same specific action carried out by the query person (e.g., fighting or bending down). This could be combined with pre-defined ones or even refine the search to detect chains of events in the previous queries.

In this paper we propose an unsupervised human action and appearance retrieval system based on an adaptive search algorithm, which focuses the analysis on relevant frames; a pedestrian detector, which selects windows containing people in these relevant frames; and a matching algorithm based on appearance and action features. The experimental results demonstrate the potential of the proposed system in a realistic surveillance video dataset that contains people performing different actions and appearing multiple times in the sequence. The retrieval process is illustrated in Fig. 1. The contributions of our work are two-fold. First, the system performs unsupervised coupled action and appearance retrieval avoiding a training stage, which has the potential of adapting to new unseen actions (i.e., no fixed set is used) performed by a given individual (i.e., no gallery is used). Second, the system provides a coupled action-appearance retrieval, which focuses the search on an individual performing an specific action as defined by the query example. Up to our knowledge, both contributions are novel in the literature.

The paper is structured as follows. Section II describes the existing work in the related topics tackled in this paper. In Section III we describe the system, specifically background subtraction, object detection, and the features and matching algorithm used in the retrieval system. Section IV presents the experimental results. Conclusions and future work are outlined in Section V.

II. Related Work

The proposed system is related to three different topics: action recognition, person re-identification and surveillance image retrieval as the target application. Although not all the papers are focused on surveillance, they are still relevant to our work. For comprehensive surveys on these topics, we refer the reader to [4], [5].

Action recognition has been researched for more than a decade, being specially relevant the works in classification. Spatial-temporal features are usually exploited to represent actions, in [6] combined with SVM and in [7] with a bag-of-words. More recent proposals consist in LDA-based action learning [3] or hierarchical representations [8]. The most extended approach is to train the models using training examples from a dataset. In our case, the search is unsupervised and does not use any training set to construct the action models. This relates to the one-example action classification approach in [9], but in our case we use a simpler representation combined with appearance.

Different approaches have been used for re-identification. Some examples are the Bayesian model of the color in three regions and the last seen locations used in [10]; color and texture cues in six horizontal regions which are fed to Support Vector Ranking in [11]; or color histograms, covariance matrix and SIFT features in [12]. A two-step ranking algorithm based on a generative model in proposed in [13]. It is based on region covariance descriptors and a boosting feature selection based on Haar-like and covariance features. A 3-stage system based on implicit shape model and SIFT features is used in [14]. In our case, we make use of a similar color representation as the used in [11].

Regarding retrieval systems in surveillance, [15] presents a detection-tracking-based human action classifier that uses CNN and SVM based on bag-of-words to classify the tracked people according to three actions of interest. Textual queries have also been proposed to retrieve specific individuals based on facial attributes [16]. Moreover, [12] presents a SQL-like language aimed at retrieving and indexing surveillance videos. This system resembles our proposal in the sense that one of the possible functionalities is to restrict the query to objects and events. However, in their case the events are not unsupervised and specific to the person but fixed and related to the context (position), e.g., entering a given region of interest.

III. Unsupervised Ranking Action Retrieval

The proposed system retrieves instances of people according to action and/or appearance criteria based on a query provided by an operator. The queries are in the form of a window of interest in an image along with a search criterion. This means that the system can search a video for a specific person performing the same action/appearance as the query one. As explained in the previous section, such a retrieval is useful to focus the investigation efforts on specific places and times, detect similar actions of a given person, and potentially extract chains of events in forensic video surveillance systems.

First, an operator selects a query window q either by explicitly defining its position and size or selecting one from a set of detected ones. Then, different action and appearance descriptors are extracted from this window depending on the criterion of the retrieval c ($c = act$ for action, $c = app$ for appearance or $c = act + app$ for both) to compute the representation q_{act}, q_{app} or $q_{act+app}$, respectively. Then, the system analyzes a given video to find matches with q. An adaptive search algorithm, which makes use of background subtraction masks, processes the video just focusing on the frames that present relevant changes with respect to the previous ones. In this way, a lot of computational time is saved and only representative frames are selected. For each analyzed frame, a human detector algorithm detects people in the foreground regions provided by the background subtraction algorithm. It is worth to highlight that no tracker is used to add temporal coherence to the detections, but the system is based on single-frame detections and the contents of the windows defined by these detections in the following frames. The advantage is that the search process is accelerated since many frames can skip the detection and tracking process. Furthermore, in this way the matching is independent from the possible errors of the tracker. Then, action and appearance descriptors are also extracted from each detected window $s \in S$, being S the set of all detected windows, depending on the retrieval criterion, getting s_{act}, s_{app} or $s_{act+app}$. Finally, a matching score is computed for each s according to the probability to match q and a retrieval criterion R. This score is used to sort the retrieved windows to be output.

Next we describe the three components of the search system.

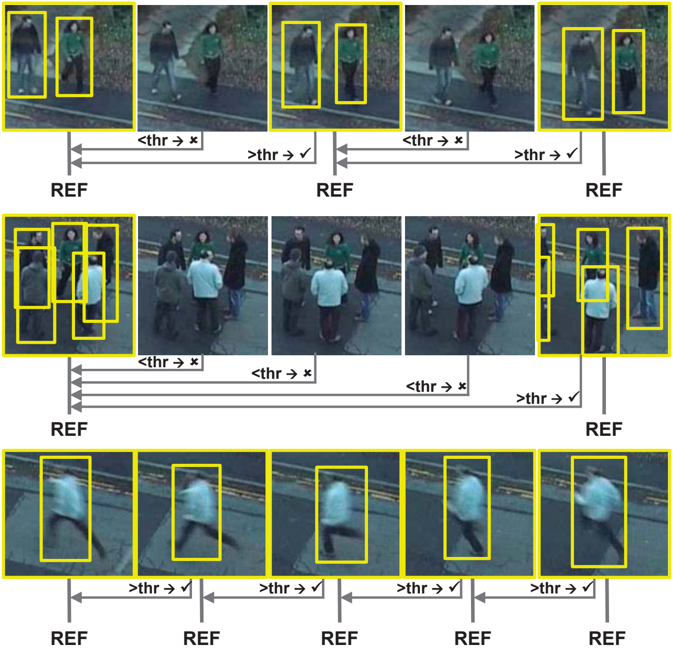

Fig. 2. Adaptive Search Algorithm. (Top) People walking present moderate scene changes, so the scanning only selects the relevant frames. (Middle) People standing present minor scene changes so the scanning process is relaxed. (Bottom) Actions such as running or vehicles crossing are densely scanned.

A. *Adaptive Search Algorithm*

The processing starts by a background subtraction algorithm that filters the non foreground regions of the scene. We make use of an adaptive Gaussian Mixture Model algorithm proposed by Zivikovic *et al.* [17]. The algorithm adjusts to changing backgrounds while detecting foreground pixels by selecting the components of the mixture model used in each pixel. Given that surveillance scenes can have long-term static objects (e.g., standing pedestrians) that are still focus of the system, the foreground mask must be computed for every analyzed frame (i.e., one out of n) for the test video in advance. Otherwise, these targets would be transparent to the search. In addition, the background model update parameter α must be high so that these objects are not rapidly integrated in the background model (see Sect. IV).

Once the background subtraction has been performed, a frame-based search is performed to detect each window to be matched to the query. The naïve search approach is to sample every frame. In this case, retrieving from a one hour video at 25fps represents scanning 90 000 frames. This poses two problems. First, the computational time to process the video is high. Second, such a dense search outputs many redundant detections (e.g., standing people in the same position for several seconds). If the sampling is relaxed to one every 5 seconds, the number of scanned frames is reduced to 18 000. However, in this case the risk is to miss fast actions, e.g., a person running can appear and disappear between these 5 seconds.

In order to solve these problems, an adaptive search algorithm based on background subtraction masks difference is proposed (Figure 2). Given an input frame, we compare it with the reference one (in the first frame, it is initialized blank) by computing the difference of their foreground masks. If the difference is higher than a threshold, it means that sufficient changes have occurred in the scene, so the frame is analyzed and set as the reference frame. Otherwise, the frame is discarded. The algorithm also includes a parameter defining the maximum number of skipped frames. This avoids skipping frames lacking of large displacements but containing local movement, e.g., people fighting. Figure 2 illustrates the adaptation of the search according to different actions. As can be seen, clips with standing people are only analyzed when there are sufficient changes in the scene, while running people are analyzed at almost every frame. With this algorithm the number of analyzed frames is not fixed but depends on the content of the video. As an example, for the test video used in Sect. IV, the number of analyzed frames is around 3000/hour.

B. *Pedestrian Detector*

For each selected frame with its corresponding foreground mask, a multi-scale sliding window selects the windows containing foreground pixels to be classified as pedestrian or non-pedestrian. We use the Integral Channel Features classifier by Dollár *et al.*. [2]. This classifier first extracts local features of the scan window from multiple cues: color, Gabor filters, Difference of Gaussians, Canny edges, gradient orientation and magnitude and binary thresholding. The most discriminative features are selected by a soft-cascade AdaBoost, which is also used to classify new samples with the learned model.

The multi-scale candidate window selection produces many overlapping hits around a pedestrian. In order to group them and discard redundant windows, we employ the pairwise non-maximum-suppression algorithm also proposed in [2].

C. *Action and Appearance Features*

Both q and s in an analyzed frame are represented by action and/or appearance descriptors (Figure 3 illustrates them). The action descriptor consists in a histogram of the optical flow in the detected window. First, the optical flow map is computed in the next 5 frames from the analyzed frame using pyramid KLT Feature Tracker [18]. Then, a histogram of the flow magnitude consisting of 2 regions, top and bottom, for each of the next 5 frames; and the variance of the unsigned flow orientation in the whole window through the same 5 frames are computed. Optical flow histograms have previously been used for action classification [19] by exploiting the orientation information. In our case, we discard the orientation in order to make the model independent from the direction and viewpoint of the person (e.g., a person walking from left to right or from top to bottom of the image should have the same representation). Regarding the appearance descriptor, we use different color spaces, which is inspired on re-identification methods like [11]. We compute the mean of R, G, B, H, S, Y, Cb and Cr color channels of the same two regions used in the action feature, top and bottom, but only in a single frame. Contrary to [11], we omit the texture information (Gabor or LBP in [11]) given that it is often not so distinctive in the type of videos we focus (i.e., low-resolution surveillance). In addition, while in [11] 6 horizontal stripes are used, in our case we only use two stripes (regions) given that the detection windows can be incorrectly aligned with the

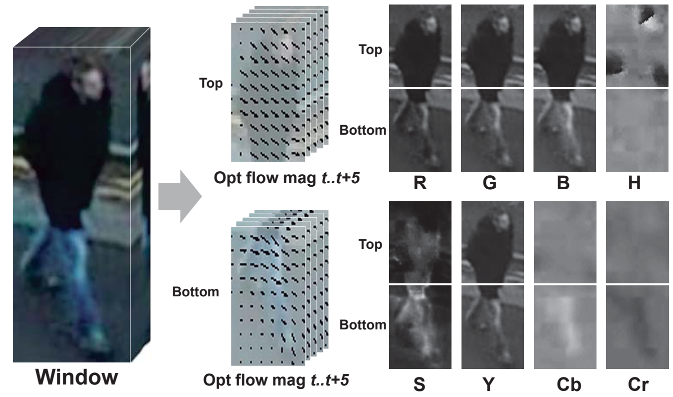

Fig. 3. Action and color representation of the query and detections. (Left) Original window. (Middle) The action representation uses a histogram of the optical flow magnitude in two regions along several frames. (Right) Appearance uses Color components in several spaces in two different regions of the window.

actual person produced by the detector and non-maximum-suppression output. For example, if the detection is slightly larger or shifted up or down with respect to the model, the representation in [11] would fail.

Once the descriptors have been detailed, the retrieval procedure is formulated as follows. Let q and s the query sample and searched detection, respectively,

$$P(r=1|s,q,c) \propto P(s|q,c,r=1)P(r=1|q,c) \qquad (1)$$

is the probability that s is relevant ($r=1$) under a retrieval criterion c, assuming independence between the variables, which happens if the surveillance video is sufficiently big. Furthermore, we assume equiprobable actions and appearances in the video, so we set $P(r=1|q,c)$ to a scaling factor. Depending on the context application, this probability could be indeed computed with training data. $P(s|q,c,r=1)$ depends on c, so it can be either $P(s_{act}|q_{act},c=act,r=1)$, $P(s_{app}|q_{app},c=app,r=1)$, or the product of both if $c=act+app$, and in that case both factors can be assumed independent. Finally, $P(s_{act}|q_{act},c=act,r=1) \sim \mathcal{N}(q,\Sigma_{act})$ and $P(s_{app}|q_{app},c=app,r=1) \sim \mathcal{N}(Q,\Sigma_{app})$.

IV. Experimental Results

This section evaluates the performance of the proposed system. We first introduce the dataset and protocol employed to evaluate the system. Then we describe the details of the system setup. Finally, we present the results.

A. Dataset and Protocol

In order to assess the performance of the proposal, a dataset containing three specific components is required. It must contain 1) different people appearing several times along the video, 2) performing different actions and 3) in a surveillance context. There are many public video datasets available for action classification (KTH, Weizmann), Internet video (YouTube Action, ASLAN), and surveillance (CAVIAR, PETS2009, VIRAT). Unfortunately, the number existing datasets suiting the aforementioned requirements is limited: either the number of actions performed by people in surveillance videos is limited to walking (e.g., VIRAT or PETS) or they are short actor-clipped non-surveillance sequences (e.g., KTH or Weizmann). On the contrary, BEHAVE dataset [20] adjusts well to our needs: it is a surveillance video with people performing different actions alone or in groups and re-appearing multiple times. This dataset was originally aimed at evaluating interaction recognition between people in groups. Several actors and non-actors perform several actions (walking, running, fighting and standing) during a 52 minutes video recorded from a regular camera (640×480 pixels at 25 fps) placed on a window facing to the street in the University of Edinburgh.

We have selected 40 query windows (all them coming from the detector output) from the BEHAVE video, 10 for each of the existing actions in the dataset (standing, walking, running and fighting). Each of these windows and the test video are fed to the system, which outputs a ranked list of retrieved query matches according to action, appearance or both. During the search, we omit one minute before and after the query time point to make the retrieved results independent from the query. From each output ranked list we compute the precision at several operation points, namely P@1, P@5, P@10 and P@20, where P@n computes the number of relevant matches over the first n matches.

B. System Setup

The background subtraction is based on bgslibrary by A. Sobral [21]. We set $alpha = 0.0001$, which results in a slow model update background subtraction that allows a pedestrian to be stopped during 40 seconds (at 25 fps) without integrating him/her into the background. The detection part makes use of Dollár's Toolbox [2], using the model trained with INRIA dataset and setting the scanning stride 4, the number of scales per octave to 16 and restricting the minimum detections to 300 pixels high. The so-called Deformable Parts Model detector has also been tested, but the performance is much lower, probably because Integral Channel Features deal better with the blurry and less contrasted BEHAVE images. Regarding the action/appearance features parameters, the action description uses 4 optical flow bins, and the covariance matrices Σ_{act} and Σ_{app} are diagonal scalar matrices experimentally set to $\lambda = 0.5$ and $\lambda = 0.7$, respectively.

C. Results

Table IV shows the quantitative performance of the system on BEHAVE dataset. Each cell is the average of 10 query examples of a specific action. As can be seen, the appearance retrieval performance is high in general. The worst cases (people running) output 13.6 correct matches in the first 20 results, while in the best cases (standing) all the 20 examples are correct. In the case of the action retrieval, the performance is lower, specially in the case of more random actions such as running or fighting, in which the precision is around 32-40%. The explanation to this decrease lays in the fact that the search query model is constructed on a single example, not an action training set containing many possible variations. Hence, if the query contains a not so representative example of the action process (e.g., a person fighting is selected in a frame in which he/she does not punch nor kick but defends or

	P@1			P@5			P@10			P@20		
	App.	Act.	App+Act	App.	Act.	App+Act	App.	Act.	App+Act	App.	Act.	App+Act
Standing	100%	100%	90%	98%	100%	86%	94%	100%	86%	94%	100%	84%
Walking	90%	60%	60%	84%	82%	62%	85%	83%	63%	84%	82%	60%
Running	80%	40%	30%	80%	32%	24%	76%	34%	27%	68%	35%	28%
Fighting	90%	30%	30%	92%	42%	36%	91%	40%	28%	86%	38%	25%

TABLE I. PRECISION RESULTS GROUPED BY DIFFERENT ACTIONS.

steps back, which happens in the selected queries), the results are noisier. In the worst cases, 7 of the 20 first retrieved results will match the searched action, which is still satisfactory for many surveilance purposes. The combined retriever presents lower performance than the individual retrievals. It can be clearly seen that the combined performance corresponds to the product of the individual retrieval performances given that its probability is computed as the product of the individual probabilities. In the worst cases, fighting and running, around a 30% of the retrieved results are relevant in average. In a realistic forensic environment, this means that when a search for a specific person in a fight, 5 of the first 20 results will contain this same person fighting. Fig. 4 shows some qualitative results that illustrate the performance of the system on two query clips and the three retrieval criteria.

Notice that no baseline has been used in the experiments since the proposed combined retrieval is the main novelty of the paper. Accordingly, we prefer to avoid naive approaches such as simpler representations as they would not vary our results and conclusions. On the other hand, it is worth to point out that the random results for such a system would be around 5%, given that there are 5 individuals who perform 4 different main actions. This number could even be lower given that not only these 5 actors appear in the whole video and that actions are inbalanced (there are more people walking than running).

In general, the incorrect matches correspond to three different causes. Firstly, false positives from the detector, which seldomly occur, could be partially avoided with improved background substraction and classifier. Secondly, wrong action retrieval is related to the difficulty of constructing a representative model from a single example, which could be improved by exploiting some prior-knowledge based on training or using more temporal information (currently less than 1 second is used in the representation). Finally, the incorrect appearance matching may happen with people wearing similar clothing, which could be amended with a richer re-identification algorithm and more precise detections (some of them are slightly shifted).

V. Conclusions

We have presented a human action and appearance surveillance image retrieval system that makes use of an adaptive search algorithm, a person detector, and optical flow and color features to represent the detections. With this approach, the number or types of actions and individuals in the system are not restricted to a specific set but new videos can be searched without additional training or annotating effort. The results are satisfactory in a surveillance scenario containing people performing different actions and appearing different times.

The future work will focus on incorporating more complex representations such as spatial-temporal bag-of-words [7], shape-based features [14] and single-example action models [9], which can potentially improve the retrieval rates in the different search criteria. Text-based queries will be also researched as a complementary retrieval interface between the operator and the system. Such textual-queries will be used either stand-alone or as a searching filter previous to the image-based query. In addition, we will explore new system functionalities such as detecting people interactions and chains of events, e.g., inferring connections between two people and specific incidents, a valuable functionality for crime investigations.

Acknowledgements

This work has been funded by a Beatriu de Pinós PostDoctoral Fellowship (EU Marie Curie Actions) of the Generalitat de Catalunya and by the Swedish Research Council under the project 2013-5473 FOVIAL.

References

[1] F. Chamasemani and L. Affendey, "Systematic review and classification on video surveillance systems," *Int. Journal Information Technology and Computer Science*, no. 7, pp. 87–102, 2013.

[2] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. British Machine Vision Conference*, London, UK, 2009.

[3] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. Journal on Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[4] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[5] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, no. 2, 2011.

[6] C. Schült, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. Int. Conf. in Pattern Recognition*, Cambridge, UK, 2004.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, Beijing, China, 2005.

[8] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[9] H. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, 2011.

[10] W. Zajdel, Z. Zivkovic, and B. Kröse, "Keeping track of humans: have I seen this person before?" in *Proc. IEEE Int. Conf. on Robotics and Automation*, Barcelona, Spain, 2005.

[11] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person reidentification by support vector ranking," in *Proc. British Machine Vision Conference*, Aberystwyth, UK, 2010.

[12] T.-L. Le, M. Thonnat, A. Boucher, and F. Brémond, "Surveillance video indexing and retrieval using object features and semantic events," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 7, pp. 1439–1476, 2009.

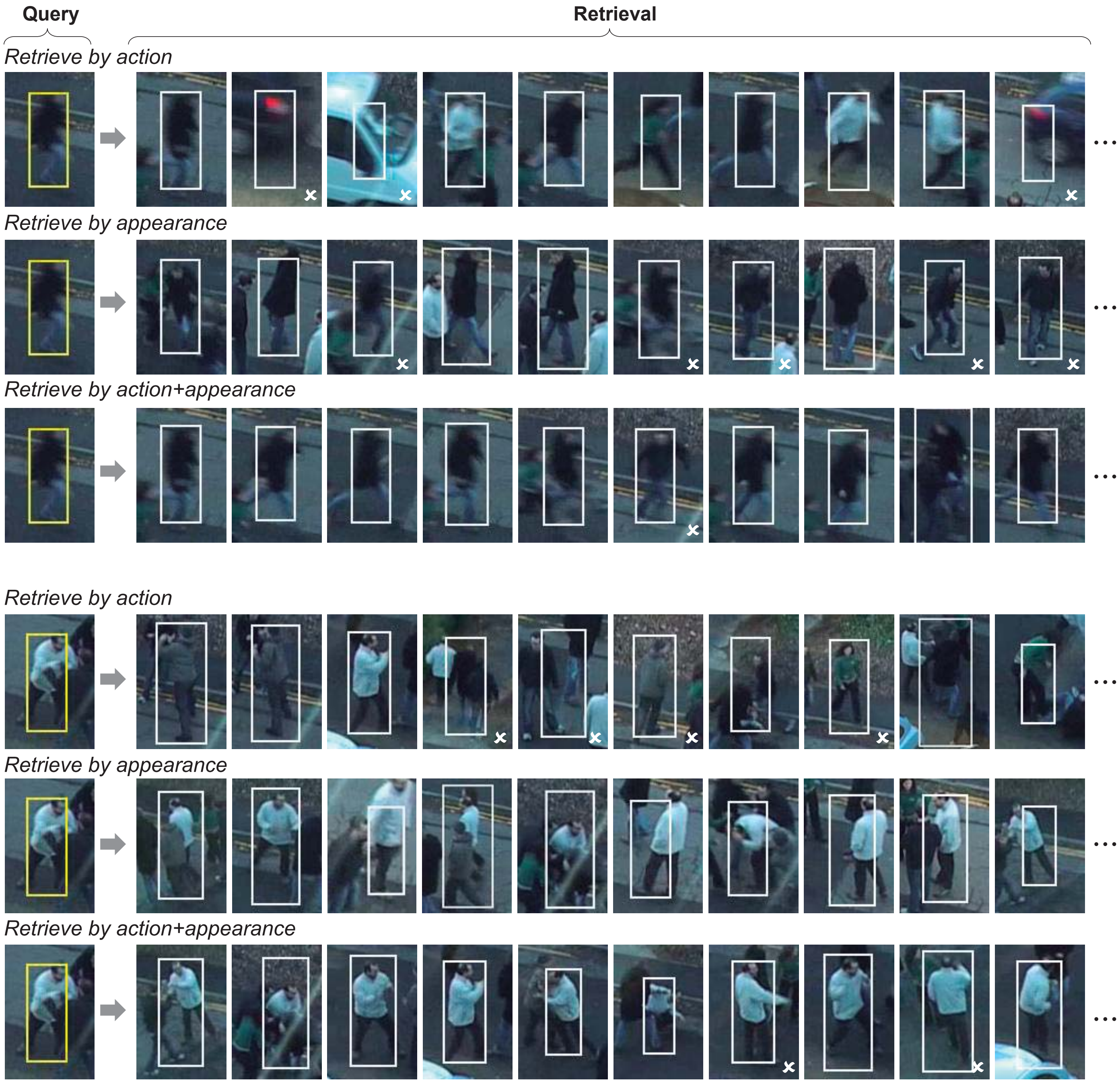

Fig. 4. Examples of queries and retrieved results. The cross marks the incorrect matches.

[13] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian Conf. on Image Analysis*, Ystad Saltsjöbad, Sweden, 2011.

[14] K. Jüngling, C. Bodensteiner, and M. Arens, "Person re-identification in multi-camera networks," in *Int. Workshop on Camera Networks and Wide-Area Scene Analysis*, Colorado Springs, CO, USA, 2011.

[15] M. Yang, S. Ji, W. Xu, F. Lv, K. Yu, Y. Gong, M. Dikmen, D. Lin, and T. Huang, "Detecting human actions in surveillance videos," in *Proc. of TRECVID 2009 Workshop*, 2009.

[16] D. Vaquero, R. S. Feris, L. Brown, and A. Hampapur, "Attribute-based people search in surveillance environments," in *Proc. Workshop on Applications of Computer Vision*, Snowbird, UT, USA, 2009.

[17] Z.Zivkovic, "Improved adaptive gausian mixture model for background subtraction," in *Proc. Int. Conf. in Pattern Recognition*, Cambridge, UK, 2004.

[18] J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[19] N. Hikizler, R. Gokberk, and P. Duygulu, "Human action recognition with line and flow histograms," in *Proc. Int. Conf. in Pattern Recognition*, Tampa, FL, USA, 2008.

[20] S. Blunsden and R. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behvior classification," *The Annals of the BMVA*, vol. 2010, no. 4, pp. 1–11, 2010.

[21] A. Sobral, "BGSLibrary: An OpenCV C++ Background Subtraction Library," in *IX Workshop de Visão Computacional (WVC'2013)*, Rio de Janeiro, Brazil, Jun 2013.