Multimedia Surveillance: Content-based Retrieval with Multicamera People Tracking

Simone Calderara
D.I.I. - University of Modena
and Reggio Emilia
Via Vignolese 905/b
Modena, Italy

Rita Cucchiara
D.I.I. - University of Modena
and Reggio Emilia
Via Vignolese 905/b
Modena, Italy

Andrea Prati
D.I.S.M.I. - University of
Modena and Reggio Emilia
Via Amendola 2 - Pd. Morselli
Reggio Emilia, Italy

ABSTRACT

Multimedia surveillance relates to the exploitation of multimedia tools for retrieving information from surveillance data, for emerging applications such as video post-analysis for forensic purposes. Searching for all the sequences in which a certain person was present is a typical query that is carried out by means of example images. Unfortunately, surveillance cameras often have low resolution, making retrieval based on appearance difficult. This paper proposes to exploit a two-step retrieval process that merges similarity-based retrieval with multicamera tracking-based retrieval able to create consistent traces of a person from different views and, thus, different resolutions. A mixture model is used to summarize these traces into a single prototype on which retrieval is performed. Experimental results demonstrate the accuracy of the retrieval process also in the case of varying illumination conditions.

Categories and Subject Descriptors

I.4.8 [**Computing Methodologies**]: Image Processing and Computer Vision—*Scene Analysis*

General Terms

Design,Algorithms,Security

Keywords

multimedia surveillance systems, people detection and tracking, consistent labeling, multiple cameras

1. INTRODUCTION

Two requirements become crucial in the new generations of surveillance systems: to be distributed and to collect multimedia surveillance data. The first requirement answers to the need of covering wide areas. Multiple fixed cameras with partially overlapped FoVs (Field of Views) are often a good solution; in addition, they help in solving occlusions by providing different viewpoints. The second requirement, instead, is driven by emerging applications in video surveillance, such as video post-analysis for forensic purposes, retrieval of events, people, or faces from videos, and comparison of multimedia sources. All these applications require to extract semantics from the video. While at event level the useful semantics heavily depends on the application, at the object level, most of the time, it is represented by the appearance of the moving objects. In video surveillance, semantics mainly concerns people's appearance and motion. Having more cameras that look at the same scene enables to collect the appearance of the same object/person from different viewpoints and at different resolutions. A new term that considers both these two requirements is "multimedia surveillance". The adjective "multimedia" is normally referred to systems and services conceived for human end-users for accessing and using multimedia data, streams, content, and interfaces. Combined with "surveillance", the term "multimedia" is not limited to define a system providing outputs in a multimedia format, but indicates a system aiming at collecting, processing in real time, correlating and handling multimedia data coming from different sources [7].

Thus, a key aspect of new multimedia surveillance systems is to put together a plethora of camera systems. Incoming video streams are, in many practical scenarios, watched by human operators and possibly stored in large repositories on central servers. Manual textual annotation is often provided to allow successive keyword-based information retrieval. A video DBMS can be created to allow standard similarity-based queries exploiting keyframe similarity. This approach is typical of *content-based video retrieval* (CBVR) systems as proposed, for instance, in [1].

With these premises, two important tasks must be accomplished: *people detection*, allowing the system to extract and log the person's appearance in the repository; *people tracking*, allowing to correlate appearance of the same person along time and, thus, to retrieve information on his path or behavior. Moreover, person's identity must be preserved also when he moves from the FoV of a camera to another. This is achieved by exploiting *consistent labeling* [12, 5, 3] that aims to coherently assign a label to all the instances of the same person on different views.

Besides the beneficial effect of keeping a person tracked on a wider area, consistent labeling also provides different images of the same person, possibly acquired at different resolutions. This feature is very useful for CBVR systems

VSSN'06, October 27, 2006, Santa Barbara, California, USA.

because usually surveillance cameras are at low resolution and placed at high positions, resulting in a person being composed of too few pixels for an effective retrieval based on pixel appearance. If a camera with more zoomed FoV is available, its view of the person can be correlated (by means of consistent labeling) to the other views, and can be used for a more effective CBVR. In other words, this approach provides multiple levels of resolution of the person trying to obtain a good area coverage and, at the same time, highly-defined captures of the person's appearance.

The research activity in distributed and multimedia surveillance systems is relatively recent; a pioneer project that defined a cooperative multisensor architecture was VSAM [6]. The review of Valera and Velastin in [14] gives a panorama of the so-called "intelligent distributed surveillance systems" typically oriented to handle multiple cameras in large environments, but no mention of multimedia is reported. In [2] Brooks et al. defined a sensor network architecture for tracking and classification. The NeST [11] architecture was proposed to enlarge the view to the campus of UCSD with many distributed cameras. The WISE architecture [13] is composed of wireless nodes and PTZ cameras and provides active surveillance of moving objects predicting their motion.

2. MULTICAMERA MULTIMEDIA SURVEILLANCE SYSTEM

As stated in the previous section, multimedia surveillance systems need to store semantics at object level by automatically annotating the person's appearance. Given a multi-camera system, classical approaches treat each camera system as independent (Fig. 1). An algorithm for people detection and tracking runs on each camera and extracts what we called *single camera appearance trace* (or $SCAT$ in short). $SCAT$ of the person P on camera C_i is composed of all the past *person's appearance* (PA) of P at instant time t: $SCAT_i^P = \{PA_i^P(t)|t = 1, ..., N_i^P\}$, where t represents the samples in time in which the person P was visible from the camera C_i and N_i^P is the total number of frames in which he was visible. As segmentation algorithm we use the background suppression method described in [8] that allows to extract the PA composed of the appearance corresponding to the mask of the person's shape. Tracking, instead, is performed by using the probabilistic tracking based on appearance reported in [9].

Our proposal exploits consistent labeling to combine the traces of the same person provided by different camera systems in order to create a *multicamera appearance trace* (or $MCAT$ in short) (Fig. 2). $MCAT$ for a person P is composed of all the $SCAT_i^P$ for any camera C_i in which, at the current moment, the person P has been detected at least for one frame, obtained by means of the consistent labeling described in the following section.

2.1 MCAT creation with consistent labeling

The proposed consistent labeling approach employs homography and epipolar geometry to solve ambiguities in the matching of people in different synchronized camera systems. Let us consider a distributed multi-camera system composed by a generic number n of cameras $C = \{C_1, ..., C_n\}$ so that for each C^i it exists at least a camera C^j whose FoV is overlapped to that of C^i.

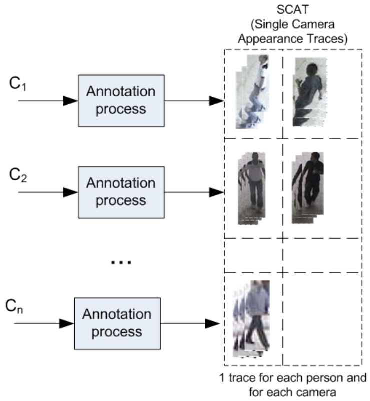


Figure 1: Standard annotation from multiple cameras.

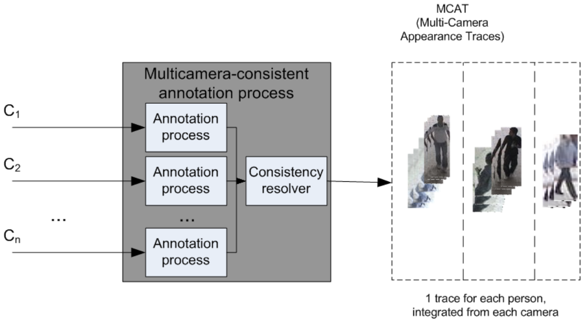


Figure 2: Proposed multicamera annotation.

As a first stage, offline computation of the homographic transformation (on the ground plane $z = 0$) and epipolar constraints is performed. This approach does not require complete camera calibration, but exploits the information computed from a video containing a single person moving freely in the scene to extract the the *Entry Edges of Field of View* and compute the homographic transformation on the ground plane [5].

Moreover, given a 3D plane Π and its projections on camera C^i and C^j (π^i and π^j, respectively), the relation between them is pointed out by the homographic matrix H. The parallax property of projective images is exploited to compute epipole location using a single plane. Given a 3D-point M_k not laying on the plane Π, and its projection $\mathbf{m_k^i}$ on C^i, it is possible to find two true correspondences in the image plane of C^j. The line computed from these points must be an epipolar line since it passes through the images of the same point of image plane I^i. Given at least two lines, the epipole can be located as the intersection of these lines and the fundamental matrix can be computed [4].

After this offline phase, the system is ready to maintain system's consistency when camera handoff occurs. To ensure system consistency we define a *maximum-a-posteriori* (MAP) estimator choosing the most probable label configuration to be assigned to new object. From this perspective, for an object appeared on camera C^l at handoff time and identified by *new*, a hypotheses' space Γ must be created for each overlapped camera, considering all the possible label assignments. Assuming that FoV of camera C^i is overlapped to that of camera C^l, hypotheses' space $\Gamma^{l,i}(\tau^l_{new})$ consists of all the possible combinations of candidate objects.

For each single hypothesis in hypotheses' space $\Gamma^{l,i}(\tau^l_{new})$, posterior probability is evaluated exploiting the Bayes rule where the hypothesis itself is considered as a single partition of the full hypotheses' space. Prior probabilities are computed by warping the lower support point of each candidate object in the image plane of the other camera. A hypothesis consisting of a single person should have higher prior probability if the warped lower support point is far enough from the other objects' support points. On the contrary, a hypothesis consisting of two or more objects (i.e., a possible group) will gain higher prior if the objects composing it will result close each other after the warping, and, at the same time, the whole group is far from other objects.

Likelihood is evaluated considering a given hypothesis on the new object's labeling and testing its fitness against current evidence. Exploiting the homography and the epipolar constraints, the principal inertial axis of the objects in each camera system can be warped on the overlapped views, supposed the objects are in the overlapping area.

Likelihood is made up of two contributions: *forward* and *backward*. The forward probability related to hypothesis $\gamma^{l,i}_k \in \Gamma^{l,i}(\tau^l_{new})$ is computed through the warping of the principal inertial axis of new object τ^l_{new} appeared in the image plane of source camera C^l at handoff-time t, while the backward probability is computed similarly to the previous contribution, except that warping source and destination camera are swapped. At the end, the maximum of the two contributions is computed and taken as likelihood value. An example of the effectiveness of the double backward/forward probability is the full characterization of groups of people, that can be realized by exploiting the forward-backward double check. Further details and experimental results can be found in [4]

2.2 System architecture

In surveillance applications, besides traditional queries such as "who was there" or similar, a significant type of queries is to retrieve the path or snapshots of a given person, starting from a selected visual object. In this *query by example* paradigm the user searches with a query image and the software finds images similar to it based on appearance.

Fig. 3 reports the retrieval process for multicamera multimedia surveillance systems. This innovative architecture divides the retrieval process in two steps: first, a *tracking-based* retrieval that uses computer-vision-based multicamera tracking techniques, and in which the selected example image $PA^P_i(t)$ is used as index for retrieving the $MCAT^P$ to which it belongs; second, a *similarity-based* retrieval in which pattern recognition techniques based on appearance is used to extract all the sequences similar to the example image. Between these two steps a further task is required. In fact, surveillance video data have two peculiarities: they provide a large amount of data, and these data exhibit a high temporal redundancy. This means that, performing the retrieval process on all the data (PA) contained in a $MCAT$, will result in high computational burden and in redundant, useless data.

Thus, the system (automatically) extracts the "best" person's appearance from the $MCAT^P$ of the given person P, selecting the camera in which the retrieval is likely to be more effective. If the query is based on the person's appearance, the best camera is that in which the person is larger, because his appearance is best represented. The system selects the camera by searching in the $MCAT^P$ the $SCAT^P_j$ with the greatest average size of the person and takes from it the $PA^P_j(t_{BEST})$ with the largest color variation. The retrieval process is then performed returning the set containing all the $MCAT$ ordered to rank the results on the basis of the similarity measure.

As a summarizing example, suppose the user at time t_0 wants to retrieve all the instances similar to the person 22 on camera 3. The system uses the appearance $PA^{22}_3(t_0)$ as index and retrieve the corresponding $MCAT^{22}$. Then, a search is performed to find the best camera (with more zoomed FoV) that results to be number 5 and, within $SCAT^{22}_5$, the $PA^{22}_5(t_{BEST})$ is selected as query object for the similarity search. The result of the retrieval process with this PA is the ordered set $\{MCAT^{22}, MCAT^{26}, ..., MCAT^{12}\}$.

3. RETRIEVAL PROCESS

During the learning phase, each $MCAT$ in the training set is processed to estimate the probability density function (PDF) of its colors (arranged in 3D RGB cube of histograms, with each dimension consisting in 256 bins) by using a *mixture of Gaussians* [10]. The process is as follows:

1. using the first PA in the $MCAT$, the ten[1] principal modes of the color histogram are extracted;

2. the Gaussians are initialized with a mean μ equal to the color corresponding to the mode and a fixed variance σ^2; weights are equally distributed for each Gaussian;

3. successive PA belonging the $MCAT$ are processed to extract again the ten main modes that are used to update the mixture; then, for each mode:

    (a) its value is checked against the mean of each Gaussian and if for none of them the difference is within 2.5 σ of the distribution, the mode generates a new Gaussian (with the same process reported above) replacing the existing Gaussian with the lowest weight;

    (b) the Mahalanobis distance is computed for every Gaussian satisfying the above-reported check, and the mode is assigned to the nearest Gaussian; the mean and the variance of the selected Gaussian are updated with the following adaptive equations:

    $$\mu_t = (1-\alpha)\,\mu_{t-1} + \alpha X_t \tag{1}$$

    $$\sigma^2_t = (1-\alpha)\,\sigma^2_{t-1} + \alpha\,(X_t - \mu_t)^T\,(X_t - \mu_t) \tag{2}$$

    where X_t is the vector with the values corresponding to the mode and α is the fixed learning factor; the weights are also updated by increasing that of the selected Gaussian and decreasing those of the other Gaussians consequently.

At the end of this learning process, ten Gaussians and the corresponding weights for each $MCAT$ are available.

[1] The choice of a mixture of ten Gaussians is motivated by the fact that a person should have at least 3-4 different modes - head, torso, harms, legs - so ten Gaussians catch all the main characteristics of its appearance

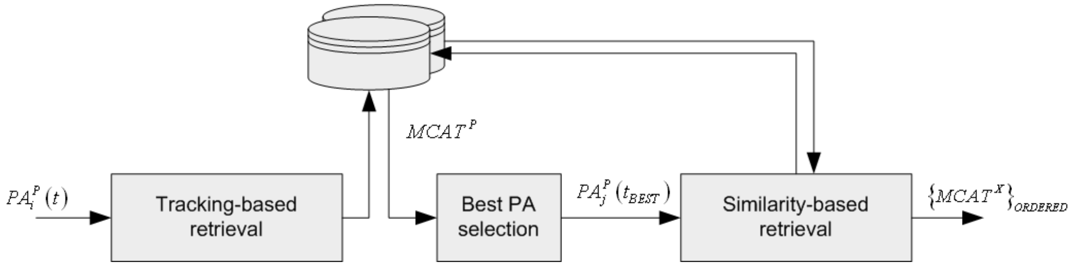


Figure 3: Retrieval process.

Queries are processed extracting, for the selected best PA (see section 2), the ten modes of the color histogram. For each $MCAT$, the ten modes X_l are compared with a subset of the ten Gaussians of the $MCAT$. The subset is built ordering the Gaussians in decreasing order respect of their weights and selecting the first n according to the following equation:

$$\sum_{i=1}^{n} \omega_i \leq TH < \sum_{i=1}^{n+1} \omega_i \tag{3}$$

The similarity measure S is achieved by computing the sum of the weights for which the difference with respect to the mean of the Gaussian is within 2.5 σ of the Gaussian:

$$S = \sum_{l=1}^{10} \sum_{i=1}^{n} b_{il} \omega_i \tag{4}$$

where:

$$b_{il} = \begin{cases} 1 & if \ \ |\mu_i - X_l| < 2.5\sigma_i \\ 0 & otherwise \end{cases} \tag{5}$$

The resulting similarity measure S is used to rank the results.

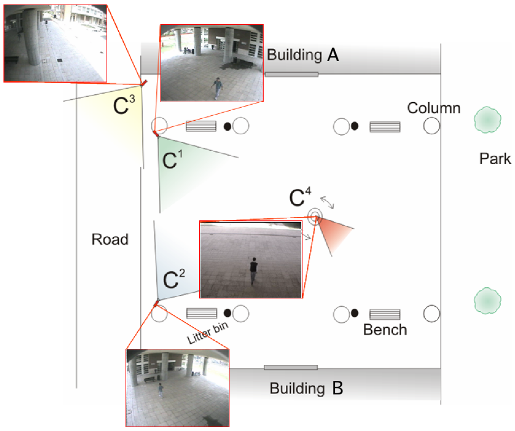


Figure 4: Map of our actual setup with some example snapshots. Cameras C^1 and C^2 have large overlapping between FoVs. In particular, both look at the space between columns, camera C^1 points at the door of building A, while camera C^2 at the door of building B.

4. EXPERIMENTAL RESULTS

Our multicamera multimedia surveillance system has been tested on a test bed with four cameras (including one PTZ camera used to acquire high resolution images), positioned as shown in Fig. 4. Tests have been performed with videos acquired during ordinary work days in different environmental conditions. The controlled area is an open space where people can enter from each side and walk freely in each direction. In acquiring the videos, no constraints have been imposed on people's trajectories or behaviors. On average, about 6000 frames have been acquired simultaneously from each camera, with about three people present in the whole scene at a time (on average). Performance of the people detection and tracking module and of the consistent labeling have been deeply reported in our previous works [8, 9, 4]. As an example of the reliability of the system also in complex situations, Fig. 5 reports examples of the segmentation, tracking and consistent labeling (same color and same id mean consistent association of label among the different views) in two-camera complex scenario, and in three-camera scenario.

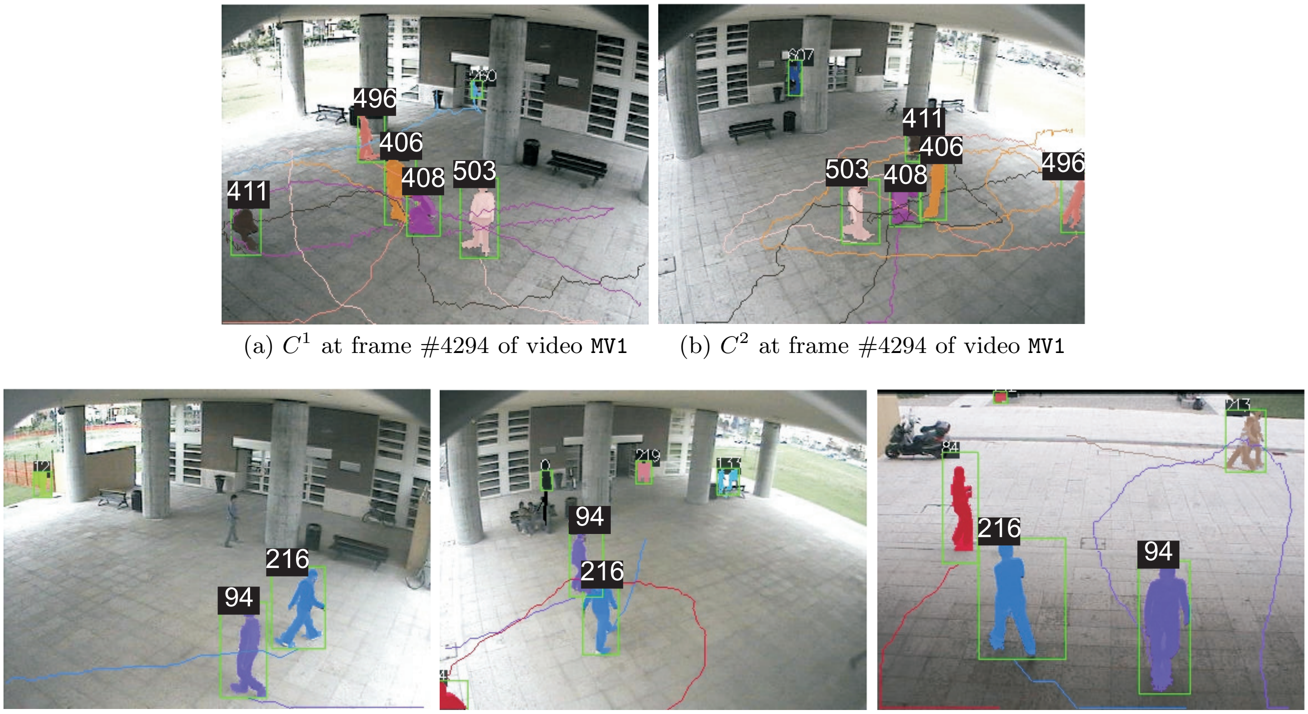
The created $MCAT$ includes approximately 700 frames per person, that guarantees to have enough data to create representative PDFs. About the 90% of the possible queries with different example images have been evaluated by computing the *recall* and *precision* achieved if only the first n positions of the ranking list are considered. The resulting graph of recall and precision in function of n is reported in Fig. 6. It is evident that considering the first 3 results we can achieve the best trade-off with both high recall (83.7%) and high precision (84.8%).

In general, the expected results are always top ranked, except for some errors due to two people very similar in appearance.

The graphs in Fig. 7 report the details on four specific queries. Similarity measure is ordered and we reported also ground truths by indicating with blue/dark bars the correct hits and with orange/light bars the non-relevant results.

5. CONCLUSIONS

The architecture for people retrieval in multicamera surveillance presented in this work embodies two main advantages with respect to conventional approaches: first, by using computer vision techniques, spatially and temporally correlated data are logically fused together; second, the proposed mixture model permits to reduce the search space, maintaining good results in terms of retrieval accuracy. Multiple camera tracking and mixture models prove to be an effective joint solution in terms of robustness, even in presence of varying conditions as commonly happens in surveillance scenarios.



(a) C^1 at frame #4294 of video `MV1` (b) C^2 at frame #4294 of video `MV1`

(c) C^1 at frame #983 of video `MV1` (d) C^2 at frame #983 of video `MV1` (e) C^3 at frame #983 of video `MV1`

Figure 5: Examples of system working on actual cases with two and three cameras.

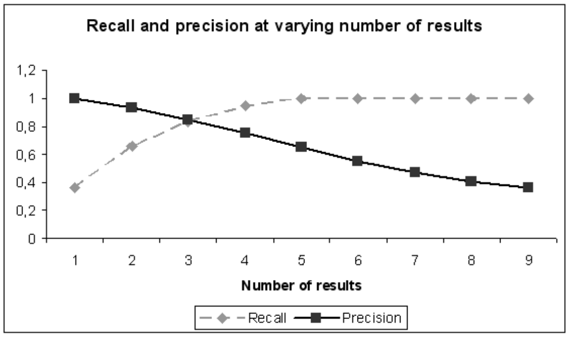


Figure 6: Recall and precision in function of the number of the positions considered in the ranking list.

6. REFERENCES

[1] G. Amato, C. Gennaro, F. Rabitti, and P. Savino. MILOS: A multimedia content management system for digital library applications. In *Proc. of ECDL - LNCS*, volume 3232, pages 14–25, September 2004.

[2] R.R. Brooks, P. Ramanathan, and A.M. Sayeed. Distributed target classification and tracking in sensor networks. *Proceedings of the IEEE*, 91(8):1163–1171, August 2003.

[3] Q. Cai and J.K. Aggarwal. Tracking human motion in a structured environment using a distributed camera system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1241–1247, November 1999.

[4] S. Calderara, R. Cucchiara, and A. Prati. Group detection at camera handoff for collecting people appearance in multi-camera systems. In *Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2006.

[5] S. Calderara, R. Vezzani, A. Prati, and R. Cucchiara. Entry edge of field of view for multi-camera tracking in distributed video surveillance. In *Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 93–98, 2005.

[6] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89:1456–1477, October 2001.

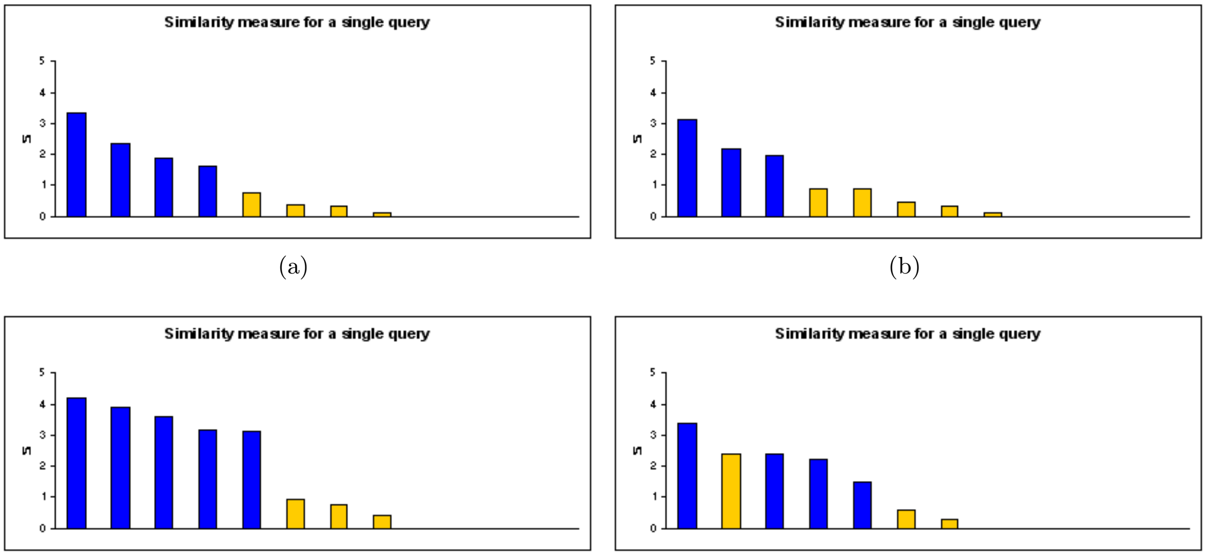
[7] R. Cucchiara. Multimedia surveillance systems. In *Proc. of Third ACM International Workshop on Video Surveillance and Sensor Networks (VSSN 2005)*, pages 3–10, November 2005.

[8] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, October 2003.

[9] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Proc. of International Conference on Pattern Recognition (ICPR 2004)*, volume 1, pages 132–135, August 2004.

[10] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (2nd ed.)*. Wiley Interscience, 2004.

[11] A. Fidaleo, H. Nguyen, and M. Trivedi. The network sensor tapestry (NeST): A privacy enhanced software architecture for interactive analysis of data in video-sensor networks. In *Proc. of ACM Workshop on*



(a) (b)

(c) (d)

Figure 7: Ordered ranking list with similarity measure: blue/dark bars indicate true positives, while orange/light bars indicate false positives.

Video Surveillance and Sensor Networks, pages 46–53, 2004.

[12] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360, October 2003.

[13] K-Y. Lam and C.K.H. Chiu. Adaptive visual object surveillance with continuously moving panning camera. In *Proc. of ACM Workshop on Video Surveillance and Sensor Networks*, pages 29–38, 2004.

[14] M. Valera Espina and S.A. Velastin. Intelligent distributed surveillance systems: A review. *IEE Proceedings - Vision, Image and Signal Processing*, 152(2):192–204, April 2005.