Object Matching in Distributed Video Surveillance Systems by LDA-Based Appearance Descriptors

Liliana Lo Presti[1], Stan Sclaroff[2], and Marco La Cascia[1]

[1] Dipartimento di Ingegneria Informatica, University of Palermo
[2] Computer Science Department, Boston University
lopresti@dinfo.unipa.it, sclaroff@cs.bu.edu, lacascia@unipa.it

Abstract. Establishing correspondences among object instances is still challenging in multi-camera surveillance systems, especially when the cameras' fields of view are non-overlapping. Spatiotemporal constraints can help in solving the correspondence problem but still leave a wide margin of uncertainty. One way to reduce this uncertainty is to use appearance information about the moving objects in the site. In this paper we present the preliminary results of a new method that can capture salient appearance characteristics at each camera node in the network. A Latent Dirichlet Allocation (LDA) model is created and maintained at each node in the camera network. Each object is encoded in terms of the LDA bag-of-words model for appearance. The encoded appearance is then used to establish probable matching across cameras. Preliminary experiments are conducted on a dataset of 20 individuals and comparison against Madden's I-MCHR is reported.

1 Introduction

In a typical video-surveillance system, the tasks of object detection and tracking across the site are crucial for enabling event retrieval and a posteriori activity analysis. Detection and tracking can be quite challenging, depending on the type of setup available. In particular, for a multi-camera system the main problem is to establish correspondences among the observations from different cameras and consistently label the objects. When the cameras have overlapping fields of view, information about the geometrical relations among the camera views can be estimated and used to establish correspondences [1,2]. In the case of disjoint views, other information about the moving objects must be used to automatically identify multiple instances of the same object [3,4].

In a distributed video-surveillance system the detected objects should be represented by compact descriptors in order to allow efficient storage of the system state and compact communication between cameras to share knowledge and apply some cooperative strategy. Compact descriptors also enable event retrieval. One possible scenario is the search of all the nodes where a certain person appears; i.e., the identification of multiple instances of the same object in different

locations and time instants. To be effective, the descriptor must also be sufficiently robust to the changes in resolution and viewpoint that occur as moving object's orientation and distance from the camera continuously change.

A number of approaches, e.g., [5,6], learn the network topology or the activity patterns in the site to predict probable correspondences among detected objects. These systems simply use the probability that an action/event will be repeated in the site during a predicted time period, and may perform not well in the case of anomalies. These approaches do not consider that multi-camera system should have a distributed knowledge of all the existent moving objects. This means that every time an object is detected for the first time, all cameras in the system are alerted about its presence in the site and wait for it reappears somewhere with characteristics quite similar to those already observed.

In a distributed system, the nodes should work like independent and autonomous agents monitoring their own FOV. Communications among nodes should be done just when objects go outside the FOV and should consist of a simple and compact appearance description related, for example, to the dress of the detected person or his identity – when possible/applicable. This approach does not require knowledge of the camera network's topology, although such information could be used to limit the number of data transmissions in the net.

Appearance is difficult to represent. Detecting objects in a well-defined way is often challenging, particularly in a cluttered environment; as a result, the information about an object is generally partial and noisy. As the object moves continuously, its appearance depends on the object's orientation with respect to the camera. Moreover, objects can move in a non-rigid way and can be self-occluded, resulting in a loss of detail.

In this paper, we present a preliminary system to model object appearance using latent features. Each object is assigned an "appearance topic" distribution from a Latent Dirichlet Allocation (LDA) model that is maintained at each camera node. The object appearance models are propagated in the network and used both to describe incoming objects and to establish correspondences. The resulting probable correspondences can be useful in constructing hypotheses about the paths of objects in the site. These hypotheses can be pruned by using the accumulated information during the life of the system or by using spatiotemporal constraints to guarantee consistency of hypotheses.

In Section 2 we describe works that use appearance information to establish object correspondences. In Sections 3 and 4 we review the LDA model and then formulate an LDA-based method that can describe objects and perform matching. Finally, in Section 5 we report experiments and compare our method's performance with that of [3].

2   Related Work

There are many kinds of approaches to establish correspondences among objects in a multi-camera system. Some of these [1,2] are based on geometrical constraints, particularly for the case of calibrated cameras and overlapping FOVs.

Other approaches instead try to find correspondences by accumulating statistics about probable associations between cameras in the network.

In [5,6,8] first the topology is estimated, then transition probabilities are used to identify where/when an object can reappear in the camera network. In this kind of system, object correspondences are strongly related to the speed at which an object moves in the site. This can result in poor performance in anomalous cases, e.g., when people do not follow the expected trajectories in the site.

Some approaches perform consistent labeling by matching features as color or texture. In [3], correspondences are found by comparing compact color histograms of the major RGB colors in the image. To make the descriptors more robust, the histogram is computed on successive frames. More details about this method will be provided in Section 5.

In [7], a content-based retrieval system for surveillance data is presented. This system looks for all the sequences of a certain person by using tracking information and appearance model similarity. By using the estimated homographic relation among the camera FOVs, consistent labeling among all the tracks is performed by assigning to each unlabeled object the label of the nearest one with the highest appearance similarity. For each observation, the ten major modes are extracted from a color histogram; then the appearance model is computed by training a mixture of Gaussian on these modes.

In [4], each object is represented as a "bag-of-visterms" where the visual words are local features. A model is created for each individual detected in the site. When a new individual is detected, classification is performed to establish a potential match with previously seen objects. Descriptors consist of 128-dimensional SIFT vectors that are quantized to form visual words using a predefined vocabulary. The vocabulary is constructed during a training step by k-means clustering and organized hierarchically so to speed up the search. Object classification is performed by an incremental version of Adaboost in which, as new data is available, classifiers are added and trained. One-vs-one SVM classifiers are added at each round; thus, the number of classifiers can grow as a quadratic function of the number of observed objects. Furthermore, the system is centralized; making it distributed requires significant communication among the nodes with a master node tasked with continuously updating the object model.

In this paper, we adopt the strategy of modeling objects as bags of words in which a latent structure of features exists and must be discovered. This structure can enable compact object description and efficient object comparison. In our distributed system, each node processes individually and autonomously the data acquired by its own camera. Communications among nodes enable knowledge sharing and are performed every time a new object exits the camera FOV. Knowledge of the camera network topology is not required. During the start-up, each node detects people and trains an LDA model. The model is then used to describe objects appearing in the FOV. Correspondences with previously seen objects are computed by comparing the stored descriptors.

3   LDA Model

Latent Dirichlet Allocation (LDA), first introduced by Blei [9], is a generative model that can be used to explain how documents are generated given a set of topics and a vocabulary of words. In the LDA model, words are the only observable variables and they implicitly reflect a latent structure, i.e., the set of T topics used to generate the document. Generally speaking, given a set of documents, the latent topic structure lies in the set of words itself. Fig. 1(a) shows the graphical model for LDA. As the figure shows, in generating the document for each word-position a topic is sampled and, conditioned from the topic, a word is selected. Each topic is chosen on the basis of the random variable θ that is sampled – for convenience – from a Dirichlet distribution $p(\theta; \alpha)$ where α is a hyperparameter. The topic z conditioned on θ and the word w conditioned on the topic and on ϕ are sampled from multinomial distributions $p(z_n|\theta)$ and $p(w_n|z_n; \phi)$ respectively. ϕ represents the word distribution over the topics. The probability of a document can be computed as

$$p(\mathbf{w}) = \int_\theta \left[\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n; \phi) p(z_n|\theta) \right] p(\theta; \alpha) \, d\theta. \tag{1}$$

There are a number of different implementations of LDA. In [9], Blei, et al. present a variational approach to approximate the topic posterior with the lower bound of a more simple and computable function. In their implementation, α and ϕ are learnt by variational inference so to maximize the log likelihood of the data.

In another approach [10] a simple modification to the model enables easier computation of the posterior (cfr. Fig. 1(b)). A Dirichlet prior is introduced on the parameter ϕ, with hyper-parameter β. Despite this modification, computation of the conditional probability $p(\mathbf{z}|\mathbf{w})$ is still unmanageable. They propose to approximate it by Gibbs sampling based on the following distribution:

$$p(z_i = j|\mathbf{z_{-i}}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha}. \tag{2}$$

This distribution represents the probability that word w_i should be assigned to topic j given all the other assignments z_{-i}. The quantities $n_{-i,j}^{(w_i)}$ and $n_{-i,j}^{(\cdot)}$ represent respectively the number of times word w_i has been already assigned to topic j and the total number of words assigned to topic j. The quantities $n_{-i,j}^{(d_i)}$ and $n_{-i}^{(d_i)}$ represent respectively the number of times the word w_i in the document d_i has been already assigned to topic j and the number of words in document d_i that are assigned to topic j. The hyperparameters α and β are computed using the method described in [10] ($\beta = 0.01$, $\alpha = 50/T$).

LDA has been applied with success in a number of computer vision scenarios. For instance, in [12] LDA is used in object segmentation and labeling for a large dataset of images. For each image in the dataset multiple segmentations are
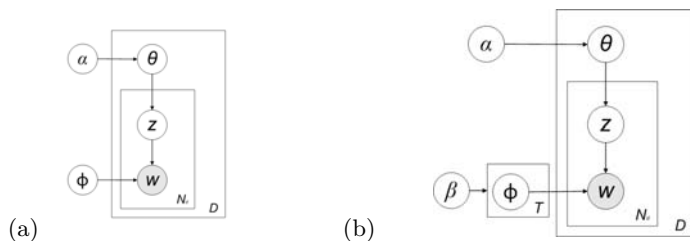
(a)   (b)

Fig. 1. Graphical models for LDA: (a) Blei's approach, (b) Griffiths' approach. In (b) a Dirichlet prior is introduced on ϕ.

obtained via different methods – these are treated as documents for LDA. An histogram of visual words (SIFT descriptors) is then computed for each segment-document. They then train an LDA model in order to discover topics in the set of documents. Segments corresponding to an object are those well explained by the discovered topics.

In other work [13], LDA is used for activity analysis in multi-camera systems. Activities are represented as motion patterns and the camera network topology is unknown. In this application, the documents are trajectory observations. Each trajectory is a set of words and a word is a tuple representing the camera, the position and the direction of motion for the observed object. Topics represent clusters of trajectories and, then, activities.

Our system detects and segments moving objects yielding their tracks in the monitored site. We define an object instance as the object segmented by background suppression from which a set of features can be extracted. Each track is a sequence of frames representing a particular instance of the object seen in different conditions, for example under different viewpoints. Indeed, during the track, objects approach or move away from the camera, are partially visible and, generally, they change their pose and/or orientation with respect to the camera.

We want to realize a system that can learn – in an unsupervised way – the appearance of the object from its track by using an LDA model. For this purpose, we treat each object instance as a document and each extracted feature as a word. In this manner, each track is just a set of documents regarding the same object. With this analogy, each object instance has been generated first choosing a mixture of features – that is a topic – then choosing a particular feature on the basis of the underlying word-topic distribution.

An example of a feature-word that can be used in our model is the pixel color. Many representations in color space can be used: HSI, RGB, normalized RGB or invariant spaces. For the sake of demonstrating the general approach, we used RGB space. Other more sophisticated features as SIFT[14] or SURF[15], could be used too but, whilst they are locally scale and rotation invariant, they are not invariant to non-rigid transformations that are commonplace in human motion patterns. Instead, using many instances of the object while it is moving should permit to capture its appearance under different viewpoints and at several distances from the camera increasing the descriptor robustness to scale changes.
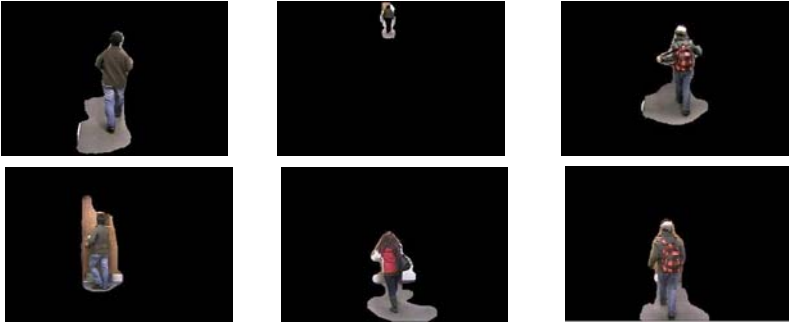
Fig. 2. Example images from the training set acquired with two different cameras. The images tend to be noisy and object resolutions differ significantly.

So, in our system each camera node computes an LDA model to capture the latent structure for the data it observes. We use the LDA method implemented by Griffiths[10]. Considering the full range of RGB colors could yield a vocabulary that is too broad; therefore, we restricted the set of words by scaling the RGB color resolution producing an 8x8x8 partition of the space. We have found in practice that this coarse uniform RGB partition tends to make the descriptor more robust to small illumination changes.

Once the model has been trained, a descriptor can be computed for every new object, based on the observed instances for the object in that camera view. As we can know the topic distribution for each object instance, we define the descriptor as the expected value of the topic distribution given the instance f. Assuming that the number of known instances is N, then our descriptor will be:

$$p(\mathbf{z}|\mathbf{Object}) = E[p(\mathbf{z}|f)] = \frac{1}{N} \sum_{i=1}^{N} p(\mathbf{z}|f_i). \tag{3}$$

Fig. 2 shows example images from the training set that we use as documents to train the LDA model. Each frame contains one object instance and has been obtained by background suppression with hysteresis thresholding. As the figure shows, the images are quite noisy and part of the background is detected.

4   Matching Object Appearances

Every time a new object is seen, a distribution of the topics must be computed for that object. Given an object and an initial topic-assignment for it, the topic distribution can be estimated by appling a Gibbs sampler to Eq. 2 where, this time, the variables $n_{-i,j}^{(w_i)}$, $n_{-i,j}^{(\cdot)}$, $n_{-i,j}^{(d_i)}$ and $n_{-i}^{(d_i)}$ are updated so to consider also the current topic-word assignment in the analyzed document.

Each node estimates its own LDA model independently from the others on different training sets, and each camera can see different object views. Therefore,

a topic association among LDA models is required in order to compare two topic distributions computed in different camera nodes. To do this, after the training is completed, the nodes propagate their own model over the camera network and then compute the topic association with each other.

Topic Association among Two LDA Models. In order to compute the topic association we consider two models, LDA1 and LDA2, with the same number of topics T and for each topic j we perform the following steps:

- we generate a document d_j by the model LDA1 using just the j-th topic; so, given $\delta_{i,j}$ is the Dirac's delta function, the topic distribution for d_j shall be

$$p_{LDA1}(z = i|d_j) = \delta_{i,j} \qquad i = 1..T \qquad (4)$$

- we compute the topic distribution of the previously generated document by using LDA2 in the manner described at the beginning of this section; i.e., we estimate $p_{LDA2}(\mathbf{z}|d_j)$.

Assuming a linear relation among the topics in the two models, performing the second step for each of the T generated documents results in a topic association matrix $\mathbf{M}$. In this manner, given a document **doc** the matrix $\mathbf{M}$ permits us to transform a generic topic distribution computed by the model LDA1 into a topic distribution $p^*_{LDA2}(\mathbf{z}|\mathbf{doc})$ valid for the model LDA2:

$$p^*_{LDA2}(\mathbf{z}|\mathbf{doc}) = \mathbf{M} \cdot p_{LDA1}(\mathbf{z}|\mathbf{doc}) \quad with \quad \mathbf{M} = \begin{bmatrix} p_{LDA2}(\mathbf{z}|d_1) \\ \vdots \\ p_{LDA2}(\mathbf{z}|d_T) \end{bmatrix}^T. \qquad (5)$$

Given this relation, for each object it is possible to compute two comparable distributions. Comparison is performed using the Jensen-Shannon (JS) divergence, which is a symmetric and normalized measure based on the Kullback Leibler divergence[11]. Defining $p = p_{LDA2}(\mathbf{z}|doc)$ and $q = p^*_{LDA2}(\mathbf{z}|doc)$, then the JS divergence

$$JS(p, q) = \frac{1}{2}\left[D(p, \frac{p+q}{2}) + D(q, \frac{p+q}{2})\right], \qquad (6)$$

where D is the Kullback Leibler divergence

$$D(p, q) = \sum_{j=1}^{T} p_j \cdot log_2(\frac{p_j}{q_j}). \qquad (7)$$

5 Experiments

To test our system we collected data using two cameras with non-overlapping fields of view. The training set comprises many different tracks of 20 different individuals acquired at approximately 15 fps, for a total of 1003 and 1433 frames

for each camera. The test set comprises many different tracks of 20 individuals acquired at different time instants, for a total of 1961 and 2068 frames per camera. None of these individuals is in the training set. Among all the possible pairs, just 10 are true matches. The number of frames per track can vary considerably, ranging from 29 to 308 frames for both the test and training set. Fig. 2 shows example images used to train the models in our experiments. As can be seen, the images tend to be rather noisy and no shadow suppression has been applied. In the images of the training and test set, the object resolution varies greatly (examples are shown in Fig. 2). In acquiring the images, the cameras' auto-focus function has been disabled as it contributes to changes in the object's appearance.

Comparison. Comparison has been conducted against the I-MCHR method of Madden [3]. As explained in Sec. 2, this method computes an incremental histogram of the object's major colors. Given the first object instance, the method computes the bin centers of the color histogram in order to obtain a rough non-uniform partition of the RGB color space; this partition is then refined by k-means clustering. Only the modes that can represent 90% of the image pixels are retained. This RGB space partition is then used to compute an incremental histogram on all the successive frames. The authors utilize a symmetric similarity measure to compare two I-MCHR descriptors by considering the probabilities of the modes with distance less than a certain threshold. Distance among clusters is computed using a normalized distance metric in the RGB space. To address the problem of illumination changes, an intensity transformation is applied separately to each image channel, thereby yielding a "controlled equalization" of the image. This transformation scales and translates the histogram modes towards the lightest part of the intensity scale. The I-MCHR method gave us an accuracy of 84% on our dataset.

Results. We tested our system with different values of the hyperparameters α and β and different numbers of topics T. In this paper we report results obtained for several values of T; β and α have been set to 0.01 and $50/T$ respectively as proposed in [10]. Figure 3(a) shows the accuracy value obtained by changing the number of topics in the range $[10; 50]$. The best results are obtained by using 15 topics, which yields an accuracy of 94%. For values of T greater than 15, the accuracy decreases.

Fig. 3(b) shows the ROC curves computed for our method and Madden's I-MCHR on the test set. As the figure shows, performance of our method is generally better than that of Madden's method. As expected, the worst results are obtained for the nosiest images, for which correspondence is particularly ambiguous (many noisy objects tend to look similar). In addition, when people dress with almost the same colors, the system cannot reliably discriminate between these individuals. In such cases, additional information could be used by the system to make a decision – for instance body proportions, distinctive gait patterns, temporal constraints, etc. The matching errors can generally be ascribed both to illumination changes and to the scaled RGB resolution of the colors we
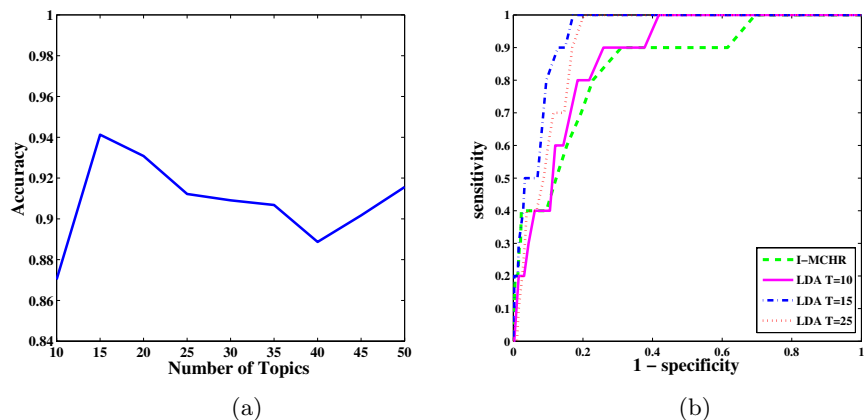
Fig. 3. (a)Accuracy of the method while the number of topics is changing; (b)ROC curves for our LDA based approach and for Madden's I-MCHR
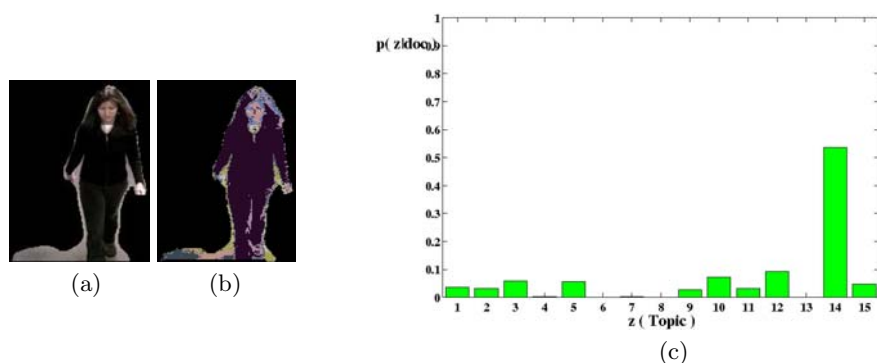

Fig. 4. Example of outputs from our method: a) the original image, b)the topic interpretation and c) the expected topic distribution (15 Topics)

used as input features to the model. Accounting for illumination variation, for instance with invariant descriptors, remains a topic for future investigation.

Nonetheless, taking into consideration that images are poorly processed and nothing is done to compensate for shadows, changes in the illumination and differences in the camera color calibrations, these preliminary results are quite promising. No doubt, further changes to the system to account for these issues can improve its overall performance.

Finally, in Figure 4 we present an example from the test set, the corresponding expected topic distribution and its topic interpretaton in false color. The latter image was obtained by associating at each pixel the most probable topic and so this image does not reflect the estimated topic distribution.

6   Conclusion and Future Work

In this paper we report the preliminary results obtained by considering an object as a bag of words and using a LDA model to infer the "appearance topic" distribution by which the object has been generated. This distribution is used to describe the object and also to establish probable correspondences among objects moving within the camera network.

Our LDA-based method performs better than Madden's I-MCHR in our experiments. Based on the preliminary study, we believe that the LDA model for appearance is promising. The formulation can be extended to include other features that describe an object's appearance, e.g., texture. In future work we intend to investigate the performance of the method in a more complex system setup with more than two cameras. We also plan to investigate the use of appearance topic distributions within a probabilistic framework for inferring likely trajectories of objects moving within the camera network.

References

1. Khan, S., Shah, M.: Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View. IEEE PAMI 25, 1355–1360 (2003)
2. Calderara, S., Prati, A., Cucchiara, R.: HECOL: Homography and epipolar-based consistent labeling for outdoor park surveillance. Computer Vision and Image Understanding Special Issue on Intelligent Visual Surveillance, 21–42 (2008)
3. Madden, C., Cheng, E.D., Piccardi, M.: Tracking people across disjoint camera views by an illumination-tolerant appearance representation. Mach. Vision Appl. 18(3), 233–247 (2007)
4. Teixeira, L.F., Corte-Real, L.: Video object matching across multiple independent views using local descriptors and adaptive learning. Pattern Recogn. Lett. 30(2), 157–167 (2009)
5. Tieu, K., Dalley, G., Grimson, W.E.L.: Inference of non-overlapping camera network topology by measuring statistical dependence. In: IEEE Proc. of ICCV 2005, vol. 2, pp. 17–21 (2005)
6. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: IEEE Proc. of CVPR, vol. 2, pp. 205–210 (2004)
7. Calderara, S., Cucchiara, R., Prati, A.: Multimedia surveillance: content-based retrieval with multicamera people tracking. In: Proc. of the 4th ACM international workshop on Video surveillance and sensor networks, pp. 95–100 (2006)
8. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Proc. of ICCV 2003, vol. 2, pp. 952–957 (2003)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Res. 3, 993–1022 (2003)
10. Griffiths, T., Steyvers, M.: Finding scientific topics. Proc. of the National Academy of Sciences 101, 5228–5235 (2004)
11. Griffiths, T., Steyvers, M.: Probabilistic Topic Models. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) LSA: A Road to Meaning
12. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. In: IEEE Proc. of CVPR 2006, vol. 2, pp. 1605–1614 (2006)

13. Wang, X., Tieu, K., Grimson, E.L.: Correspondence-Free Activity Analysis and Scene Modeling in Multiple Camera Views. In: IEEE PAMI (accepted, 2009)
14. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004)
15. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). Comput. Vis. Image Underst. 110(3), 346–359 (2008)