Multicamera Video Summarization from Optimal Reconstruction

Carter De Leo and B.S. Manjunath

University of California, Santa Barbara

Abstract. We propose a principled approach to video summarization using optimal reconstruction as a metric to guide the creation of the summary output. The spatio-temporal video patches included in the summary are viewed as observations about the local motion of the original input video and are chosen to minimize the reconstruction error of the missing observations under a set of learned predictive models. The method is demonstrated using fixed-viewpoint video sequences and shown to generalize to multiple camera systems with disjoint views, which can share activity already summarized in one view to inform the summary of another. The results show that this approach can significantly reduce or even eliminate the inclusion of patches in the summary that contain activities from the video that are already expected based on other summary patches, leading to a more concise output.

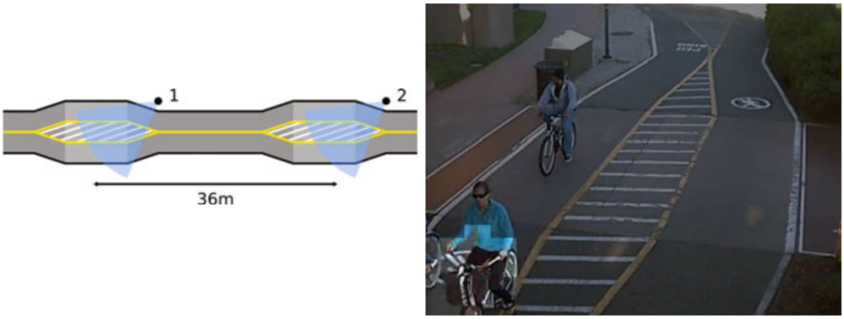
1 Introduction

Many domains, from surveillance to biology, can benefit from collecting large quantities of video data. However, long recordings over many deployed cameras can easily overwhelm a human operator's ability to review, preventing the data from being as useful as possible. In many applications with stationary cameras, much of the recorded video is uninteresting, so time spent having a human review it is wasted. Video summarization aims to highlight the most important segments of an input video, helping to focus reviewing time where it is most beneficial.

The concept of extracting the important portions of a video is not usually well defined, since importance is a subjective notion. While looking at motion or color contrast can serve as an approximation to importance, these methods take an indirect approach to the summarization problem. Instead, we propose a method that formulates the problem in a more principled way that easily generalizes to multiple cameras.

Videos from within a single camera or from close-by cameras in a network also exhibit redundancy in what they display, since activities in one region are often closely related to activities in another. For example, refer to Figure 1(a), which shows a network of two cameras positioned along a bike path. When a person leaves the view of camera 1 traveling to the right, it is expected that they will appear in camera 2 after a delay. If the delay does not greatly deviate from the average trip time observed over many people, showing the person in both views 1 and 2 is redundant; if a human observer sees the person in one view, they already

R. Koch et al. (Eds.): ACCV 2010 Workshops, Part I, LNCS 6468, pp. 94–103, 2011.



(a) (b)

Fig. 1. Layout of the two-camera network used for the experiments and a sample spatio-temporal patch drawn from camera 1 to be highlighted in the summary

have a good understanding of what happened in the other. In this case, a good summary should devote less time to the appearance of the person in one view after establishing the person's presence in the other. However, if the travel time does significantly differ from what is expected, the summary should spend extra time presenting this anomaly. This shows that a good summary would respond not just to motion, but also to whether that motion is already expected.

We view the output summary video as a set of observations on the original input video. Since the summary video is a reduced form of the input video, many possible observations are missing. The best set of observations to chose for the summary can be understood as those that, taken alone, would allow us to best reconstruct the missing data. These observations take the form of a spatio-temporal patch highlighted in the summary output, such as the example patch in Figure 1(b). Reconstruction requires a predictive model to describe how an observation at one spatio-temporal location influences the state at others, which the system can learn over local regions of the video itself since the camera viewpoint is static. This captures the intuition that if a reviewer is familiar with what normally occurs within a scene, they have effectively learned a predictive model themselves. As such, a summary consisting of the observations that give the best reconstruction of the missing data, or the rest of input video, would also give a reviewer the best mental reconstruction of what occurred.

2 Related Work

Video summarization, as well as the related problem of video anomaly detection, has been well studied in the literature, so we discuss only a subset of the past work here. Approaches tend to be divided between methods using tracked object paths and those that use features that do not rely on tracking. Systems that use tracking [1,2,3] attempt to extract the trajectories of objects of interest within a scene, then cluster those trajectories to identify outliers. Objects following unusual trajectories are then assumed to be interesting. In visually challenging scenes, however, extracting suitable object trajectories may be difficult, degrading performance.

Many other systems rely on determining the similarity between frames using other features. Examples include gradient orientations [4,5], local motion [6], and color and texture [7,8]. Another approach is to apply seam carving techniques to videos [9] to remove regions with smooth colors. While these systems can yield satisfying results, they are not directly attempting to make the most interesting or representative portions of the input video appear in the summary, instead relying on related indicators.

The system of Simakov et al.[10] does approach the summarization problem more directly by trying to choose patches of the input video to include in the summary to simultaneously maximize measurements of completeness, or how much data from the input is present in the output, and coherence, or that everything in the output was also in the input. This takes the viewpoint that a good summary is one that includes as much of the input data as possible within a constrained space without introducing artifacts, whereas our proposed approach considers a good summary as one that best allows for data missing from the summary to be inferred, thus representing the entire input.

3 Approach

Our goal is to analyze a set of input videos and determine the subset of spatio-temporal patches from them that would best summarize their contents. These can be packed into a shorter output video for a human operator to review. The first step is a scene decomposition to group camera views into regions, followed by feature clustering and region linking to cluster activities occurring in each region and determine region topology. The system learns occurrence models for the activities and then uses a genetic algorithm to seek the summary that best represents the activity sequence occurring in a region. Here, a summary refers to any selected subset of key patches. The algorithm grades the fitness of a candidate summary by finding the error of the resulting reconstruction, defined as the estimate of the complete sequence of activity labels given the subset in the candidate summary.

3.1 Scene Decomposition

Our system starts with a scene decomposition to spatially divide the input videos into regions that tend to move similarly, based on the work by Loy[11]. We follow their approach except for a change in the activity feature used. Unlike the low framerate videos presented in those experiments, the videos used here have an average frame rate around 15-20 fps. This allows the use of optical flow as the activity feature instead of the features used by Loy to accommodate low temporal resolution. We calculate the affinity matrix $\mathbf{A}$ between the 10x10 pixel, non-overlapping subblocks with sufficient activity in the video. Spectral clustering

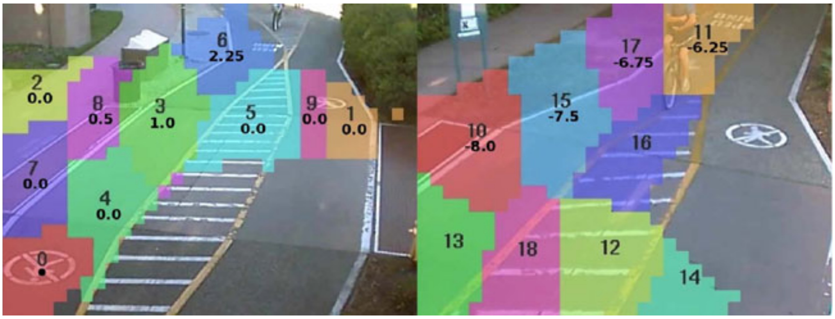


Fig. 2. Discovered links and time shifts to region 0. (Best viewed in color).

on **A** by the method presented by Zelnik-Manor[12] gives the segmented regions. An example segmentation for a video in our data set appears in Figure 2. Notice that the segmentation has separated the regions covering the bike path from the pedestrian areas on either side, giving the regions semantic meaning.

3.2 Feature Clustering and Region Linking

With the scene decomposition done, the average optical flow vector over a region can be calculated at each frame and then clustered by fitting a GMM. The number of clusters K_i for the i^{th} region is determined automatically using the Akaike information criterion[13]. Now the activity in each region can be succinctly represented by a single sequence $\mathbf{y}_i$ with $y_{i,t} \in [0, K_i)$ consisting of cluster indices over time.

This representation also allows discovery of the linkages between regions and the typical time lag between activity in one region leading to activity in another. For a proposed linkage between regions i and j for time lag τ, we can calculate the Time Delayed Mutual Information[14]:

$$\mathbf{I}_{i,j}(\tau) = \sum_{y_i} \sum_{y_j} p(y_{i,t}, y_{j,t+\tau}) \ln \left[\frac{p(y_{i,t}, y_{j,t+\tau})}{p(y_{i,t})p(y_{j,t+\tau})} \right] \tag{1}$$

The probability distributions are estimated by counting activity occurrences over the length of the videos. For each local maxima of $\mathbf{I}_{i,j}(\tau)$ for $\tau \in [-\tau_{max}, \tau_{max}]$ that exceeds a threshold I_{min}, define a link between regions i and j with a time shift of τ. We do not consider regions within a camera view differently from regions appearing in different camera views, so this linkage discovery naturally extends to a multicamera network. As an example, we use video segments collected from the two cameras shown in Figure 1(a). Figure 2 shows the resulting linkages from region 0, in the lower left corner of the first camera, to all other regions. The labels show the relative time shift τ in seconds for the link between that region and region 0; regions without labels are not connected to region 0.

3.3 Learning Occurrence Models

The system uses the set of component index sequences $\{\mathbf{y}_i \forall i\}$ to learn a set of occurrence models. For each region, estimate $p(y_{i,t})$, $p(y_{i,t+1}|y_{i,t})$, and $p(y_{j,t+\tau}|y_{i,t})$ over $(j, \tau) \in \mathcal{L}(i)$, the set of regions and corresponding time shifts that form links to region i. From these, compute the negative-log costs for assigning indices to patches:

$$\begin{aligned} c_i^p(q) &= -\ln(p(y_{i,t} = q)) \\ c_i^f(r|q) &= -\ln(p(y_{i,t+1} = r|y_{i,t} = q)) \\ c_{ij,\tau}^l(r|q) &= -\ln(p(y_{j,t+\tau} = r|y_{i,t} = q)) \quad (j, \tau) \in \mathcal{L}(i) \end{aligned} \tag{2}$$

c^p is the prior cost, or the cost of assigning component index q to a patch without knowledge of surrounding patches. c^f is the forward cost, or the cost of assigning index r to a patch when its temporal predecessor has index q. Finally, c^l is the lateral cost, or the cost of assigning index r to a patch in region j when it is linked with time shift τ to a patch in region i that has index q.

3.4 Single Region Activity Reconstruction

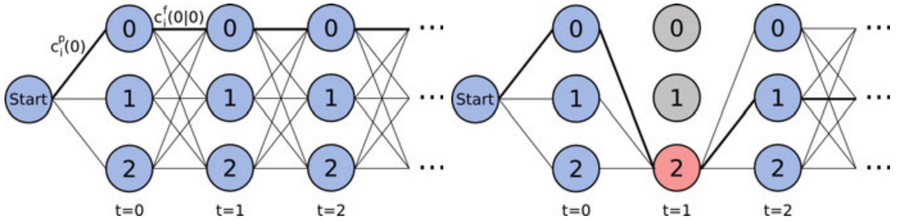
Our goal is to reconstruct an index sequence by selecting a subset of the patches from the corresponding region to include in the summary. Selected key patches in the sequence act as observed states, while the remaining patches act as missing observations. As such, the system uses a modified Viterbi algorithm with the prior and forward cost models to choose the most likely sequence of indices that explain the chosen observations. This Viterbi lattice is illustrated in Figure 3(a), where the columns correspond to steps in time and the rows correspond to the possible activity indicies for the region in the range $[0, K_i)$. The costs defined in the previous section determine the costs used to label the lattice edges. Specifically, for region i, the edge from the starting node to the node for activity q in the t=0 layer is labeled using $c_i^p(q)$. For an edge from the node for activity q to activity r in the next layer, use $c_i^f(r|q)$. In this example, the optimal path through the lattice is shown in bold.

Choosing a patch as a key patch amounts to forcing a step in the lattice to take the state seen in the input video, as in Figure 3(b). This choice updates the optimal path between states. Call $\hat{\mathbf{y}}_{i|\mathcal{P}_i}$ the reconstructed sequence after choosing to force the patches of region i in set $\mathcal{P}_i$ to their correct values. The error for this choice of key patches is:

$$\mathrm{E}_i(\mathcal{P}_i) = \sum_t ec(y_{i,t} = q, \hat{y}_{i,t|\mathcal{P}_i} = r) \tag{3}$$

$$ec(q, r) = \sqrt{\boldsymbol{\mu}_{i,q} \mathbf{C}_{i,q}^{-1} \boldsymbol{\mu}_{i,r}}$$

where the error cost ec of reconstructing a patch as having index r when it was actually q is the Mahalanobis distance from the correct GMM cluster, with mean $\boldsymbol{\mu}_{i,q}$ and covariance $\mathbf{C}_{i,q}$, to $\boldsymbol{\mu}_{i,r}$, the mean of the classified cluster.



(a) Viterbi lattice

(b) Lattice with forced step

Fig. 3. Viterbi lattice for region activity reconstruction before and after forcing. Columns represent time and rows represent cluster indices. The optimal path is shown in bold.

3.5 Key Patch Selection

To form the summary, we would ideally like to find $\mathcal{P}_i$ such that:

$$\mathcal{P}_i = \underset{\mathcal{P}'_i}{\operatorname{argmin}}\ \mathrm{E}_i(\mathcal{P}'_i) \tag{4}$$

However, there are many possible choices for $\mathcal{P}_i$; even a one minute sequence from our data set has about 1200 frames, so choosing 10% of them to include in a summary would give around 10^{168} choices. Since it is not imperative that we find the globally best $\mathcal{P}_i$ instead of a merely good one, a genetic algorithm is an appropriate way to examine such a large search space. A proposed $\mathcal{P}_i$ can be naturally represented as a binary string with length equal to the number of frames and ones in the positions corresponding to patches included in the summary, so this problem maps directly to a genetic approach. We use a modified version of the CHC algorithm[15], which stands for *cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation.* The CHC algorithm employs an aggressive search that ensures non-decreasing fitness of the best solution between generations, offset by periodic reinitialization of the population of solutions to discourage convergence on local maxima. For a proposed $\mathcal{P}_i$, we evaluate its fitness as:

$$\mathrm{F}(\mathcal{P}_i;\alpha,\beta) = \frac{\mathrm{E}_i(\emptyset) - \mathrm{E}_i(\mathcal{P}_i)}{|\mathcal{P}_i|} \cdot \exp\left(-\frac{(\alpha - |\mathcal{P}_i|)^2}{2\beta^2} \cdot 1(\mu - |\mathcal{P}_i|)\right) \tag{5}$$

This consists of two terms. The first is an efficiency term, which rates solutions higher that have achieved a large reduction in the reconstruction error per patch that it has forced. The second is a falloff term that penalizes solutions that are more concise than the target level of summarization α, but has no effect on longer solutions. Empirically, shorter solutions tend to be more efficient, so this term prevents selective pressure from creating a summary that is much more concise that the user wishes. Instead, we favor solutions that spend extra forced patches reducing the reconstruction error even modestly instead of forgoing them all together. The factor β controls how steep this penalty should be and is set such that $\sqrt{2}\beta = \alpha/10$ for all of our experiments.

3.6 Extension to Multiple Regions

If two regions i and j were found to be linked in the preceding steps, then knowing the activity present in i should also tell a human viewer something about the activity occurring in j, with some possible time shift. In our example videos, if the generated summary for region 6 establishes that a bicyclist is traveling to the left along the path, the viewer already assumes that the same bicyclist will shortly appear in region 0; if this occurs, the summary does not need to choose as many patches in region 0 to make this clear. This intuition naturally extends across cameras as well; after seeing the bicyclist leave region 0, the viewer can expect to see the same person again in region 11. If the reappearance happens close to the time shift discovered for that region link, showing that activity is largely redundant. If the actual delay differs significantly from τ, then something unusual may have happened, and the summary should spend additional summary patches illustrating this.

Formally, we can incorporate information coming from a linked region within the lattice framework by altering the transition costs for a time step using the lateral cost models learned earlier:

$$c_{i,t}(r) = c_i^f(r|y_{i,t-1}) + \sum_{(j,\tau)\in\mathcal{L}(i)} c^l_{ji,-\tau}(r|y_{j,\hat{t}-\tau}) \tag{6}$$

This represents the cost for selecting cluster label r for frame t in region i. The first term in the sum is the existing cost based on the intra-region forward model. The second term has been added to account for influence from other regions on the current region's lattice solution, based on the inter-region lateral model.

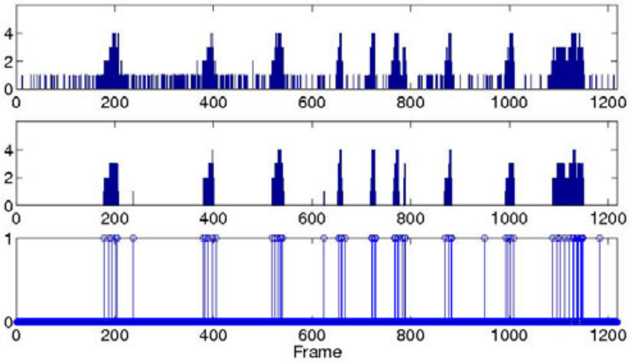


Fig. 4. Single region summarization for region 0 for a 5% target length. Top: Actual sequence. Middle: Reconstructed sequence. Bottom: Chosen summary patches.

4 Experiments

4.1 Single Region

Figure 4 shows the resulting summaries generated by the system for a one minute sample of the video corresponding to region 0 for a 5% summary length target. The top row shows the actual activity sequence for the region and the second row shows the reconstructed sequence. The third row shows spikes corresponding to the patches chosen for the summary. These are the observations in time used to generate the reconstruction. Notice that the density of the spikes is greatest where the activity indicies change, which corresponds to bicyclists moving through the region in the original video.

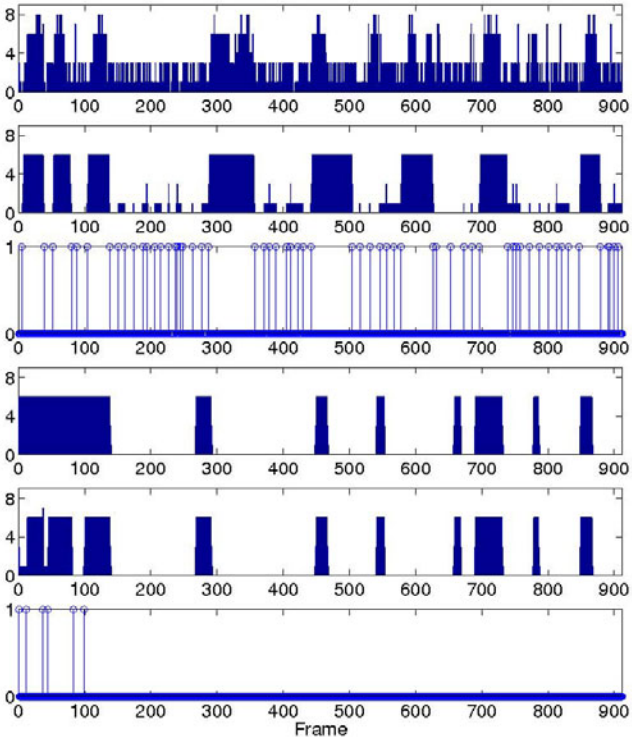


Fig. 5. Summarization of region 15 with and without information from region 0. Row 1: Actual sequence. Row 2: 5% reconstruction of region 15 in isolation. Row 3: Patches chosen for the reconstruction in the row above. Row 4: Reconstruction of region 15 incorporating information from region 0. No patches from region 15 have have been chosen. Row 5: Reconstruction of region 15 with region 0 information and choosing patches to give error equal to 5% reconstruction in isolation. Row 6: Patches chosen for the reconstruction in the row above.

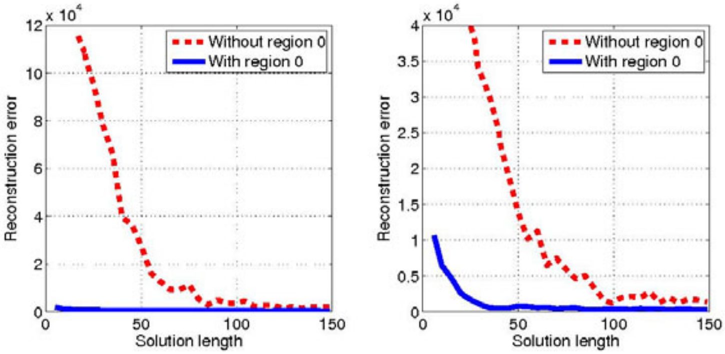


Fig. 6. Difference in reconstruction error versus summary length when excluding and including information from region 0. Left plot is for region 4, right plot is for region 15.

4.2 Multiple Regions

Figure 5 shows the effect on a neighboring region's information on the reconstruction of region 15. The first row shows the actual activity sequence for region 4. The second and third rows show the single region reconstruction of region 15 with a 5% length target and the chosen patches, as in the previous section. The fourth row shows the unforced reconstruction of region 15, or the reconstruction before any key patches have been chosen from 15 when its link to region 0 is included in the costs calculated from Equation 6. To produce this, we first generate the 5% length reconstruction for region 0 as shown in the previous section and then use the resulting $\hat{\mathbf{y}_0}$ in Equation 6 to determine the optimal path through the lattice for region 15. This shows that even if the summary did not include any patches from region 15, seeing region 0, which is in a different camera view, has already provided an idea of its activity. The fifth and sixth rows show the reconstruction and chosen patches of region 15 incorporating information from region 0 and choosing enough patches to make the total error equal to that from the reconstruction in isolation seen in the second row. Here, the system can provide observations on region 15 to correct deviations in its activity from what would be predicted by region 0. In this example, the system reaches the same total error as it did with 60 patches in isolation with only 6 patches when inter-region information is incorporated.

The benefit gained from considering information from linked regions reaches a saturation point as the algorithm includes additional patches in the summary. Figure 6 shows the total reconstruction error for regions 4 and 15 versus the summary target length, both when no inter-region information is included and when region 0 is included. Notice that using inter-region information helps provide a lower error reconstruction for a given summary length. However, since the system only needs to correct deviations from expected activity when using information from region 0's summary, it experiences less benefit by allowing a longer summary. The horizontal distance between the two curves shows the decrease in the number of frames that need to be displayed to the user after including inter-region information. Notice that for region 4, the unforced reconstruction using region 0's summary already has lower error than a reconstruction in isolation for many target lengths, so it could be excluded from the summary completely.

5 Conclusion

We proposed a technique for video summarization that takes a principled approach to creating an output summary video. By viewing the spatio-temporal patches that are retained for the output summary as observations of the local motion of the input video, our system attempts to optimally construct the summary to best allow inference of the missing input data. This allows it to choose key patches not just based on motion, but on a viewer's expectation of what motion will occur. Our results show the validity of this approach and its ability to generalize to camera networks with disjoint views by allowing motion shown in one region to inform what is shown in another, creating a more concise summary.

References

1. Wang, X., Tieu, K., Grimson, W.: Correspondence-free activity analysis and scene modeling in multiple camera views. PAMI (2009)
2. Wang, X., Ma, K., Ng, G., Grimson, W.: Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In: CVPR (2008)
3. Piciarelli, C., Micheloni, C., Foresti, G.L.: Trajectory-based anomalous event detection. IEEE Trans. Circuits Systems Vid. Tech. 18, 1544–1554 (2008)
4. Breitenstein, M., Grabner, H., Gool, L.V.: Hunting nessie – real-time abnormality detection from webcams. In: ICCV WS on VS (2009)
5. Pritch, Y., Ratovitch, S., Hendel, A., Peleg, S.: Clustered synopsis of surveillance video. In: AVSS (2009)
6. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. PAMI (2008)
7. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: CVPR (2004)
8. Zhu, X., Wu, X., Fan, J., Elmagarmid, A., Aref, W.: Exploring video content structure for hierarchical summarization. Multimedia Systems 10, 98–115 (2004)
9. Chen, B., Sen, P.: Video carving. Eurographics (2008)
10. Simakov, D., Caspi, Y., Irani, M., Shechtman, E.: Summarizing visual data using bidirectional similarity. In: CVPR (2008)
11. Loy, C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: CVPR, pp. 1988–1995 (2009)
12. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS (2004)
13. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automatic Control 19, 716–723 (1974)
14. Loy, C., Xiang, T., Gong, S.: Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In: ICCV (2009)
15. Eshelman, L.: The chc adaptive search algorithm. Foundations of Genetic Algorithms, 256–283 (1991)