Wisdom of the Crowd in Egocentric Video Curation

Yedid Hoshen* Gil Ben-Artzi* Shmuel Peleg
The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

Videos recorded by wearable egocentric cameras often suffer from quality degradations that cannot be corrected. When several wearable video cameras are viewing the same scene, it is possible to combine their multiple videos into a single high-quality video. Existing techniques select for each point in time the video having highest quality, but the highest quality video may not be relevant. E.g. the best quality video can come from a person that happens to look sideways from the main attraction.

We propose the curation of a single video stream from multiple egocentric videos by requiring that the selected video will also view the most interesting region in the scene. Importance of a region is determined by the "wisdom of the crowd", i.e. the number of cameras looking at a region. The resulting video is more interesting and of a higher quality than any individual video streams can possibly obtain. Several examples are presented demonstrating the effectiveness of this technique.

1. Introduction

The use of wearable egocentric video cameras is increasing, and one day such cameras may be used daily by many people. A notable aspect of wearable video is that popular scenes will often be observed simultaneously by multiple video cameras. For example a lecture will be simultaneously recorded by many students in the audience. This means that several similar videos of the lecture will be created. The redundancy of videos in popular scenes creates both a challenge and an opportunity for video curation and summarization techniques.

A key issue with egocentric video is shakiness and blur, as human heads perform drastic motions very frequently. Much work has been done on video stabilization ([8, 12, 6]), but even state of the art stabilization techniques frequently fail when applied to egocentric videos. Another key issue with egocentric video is that humans often perform actions that they would not wish to share in the output videos e.g. staring at their phone or out of the window. The above issues make egocentric videos of events rather poor and users are unlikely to share their videos without significant curation.

In both cases mentioned above, information is lost either by the motion of the camera or by not observing the region of interest. The most that can be hoped for by single video techniques is good stabilization and interpolation over the missing frames.

Current video stream selection techniques deal with hand-held cameras that are actively directed to point at the region of interest. Egocentric cameras do not have this luxury as they are always on and mostly do not have viewfinders. A key challenge is to classify which frames do not view the region of interest and thus should not be shown (e.g. those containing the user looking at his watch or mobile phone). Although much work has been done on determining "interestingness" ([7]), it is far from solved as it requires good high-level understanding of the scene.

We present a novel method for the curation of high-quality video from a collection of low-quality egocentric videos. We assume that the region of interest was observed by multiple egocentric cameras, each stream being shaky and occasionally observing uninteresting regions. As mentioned above, single video enhancement techniques (stabilization, deblurring, interesting frame detection) will yield good quality results for each stream only some of the time. On the other hand our approach is to combine all video streams so that high-quality results are obtained at all times.

We rely on two properties of egocentric videos:

1. Unintentional motion is uncorrelated between individual users. The periods of blurry and shaky in videos of individual users will usually occur at different times.

2. "Wisdom of the crowd" - at any given time most users look at the region of interest in the scene. This is similar to saying that procrastination is uncorrelated between different users. E.g. different users probably look at their watch or at the window at different times.

*Joint First Authors

These properties can be exploited in several ways. The first property implies that at each time interval some videos would be more stable and sharper than others. By choosing the best quality stream at each time period we obtain an output video of the highest possible quality. The second property implies that by determining which cameras look at the same region, we can detect the most popular region in the scene and can therefore reject streams looking at uninteresting regions.

The result of our method is a video containing a sequence of high quality and interesting frames. As transitions between streams usually create a sharp jump between two adjacent frames, our method also minimizes the number of transitions between streams. We also prefer to transition between streams with closely overlapping fields of view to minimize the discomfort of such transition.

The contributions of our paper are: (1) A measure for the "intrestingness" of the frames ('popularity' measure) which is based solely upon the chromo-temporal similarity between frames. Such information is available in many common scenarios resulting in a highly robust method. (2) A method for matching scene region of interest between frames across streams. Our method uses HMM and Normalized cuts. (3) An objective function based upon quality, popularity and smoothness requirements for creating a single high quality video from multiple videos solved by dynamic programming.

We begin by describing our video quality measures involving stability and sharpness of each frame (Sec. 3.1). We then describe our method for determining popular and unpopular frames (Sec. 3.2). In Sec. 3.3 a method for determining the cost of transition between two video streams based on region overlap is described. Using the above measures, an optimization problem is formulated for determining the best set of frames maximizing quality and popularity scores while minimizing sharp transitions between frames (Sec. 3.4). Several experiments are shown (Sec. 4) demonstrating the relevance and effectiveness of our method.

2. Previous Work

2.1. Single Video Enhancement

The most relevant work on video enhancement address video stabilization and deblurring.

Video stabilization is an established field of research. Early papers in video stabilization estimated 2D shifts between the image planes of subsequent frames [13]. More stable motion is obtained by either fixing or smoothing the shifts. Later papers [4] attempted to recover the 3D motion of the camera. By recovering and smoothing the camera motion in 3D, higher quality results can be obtained. Recent papers [8, 12, 6] assume that accurate recovery of 3D geometry and motion often fails. Instead they advocate combining 3D constraints with 2D image plane motion to obtain state of the art stabilization results.

Current stabilization method can perform very well for many types of videos, but are not robust for egocentric video. As noted in a recent paper [11], large and fast displacements caused by head motion are difficult to track, giving generally poor stabilization.

Motion deblurring is also a very active field. Recent papers include [10, 5]. Blind blur kernel estimation is a slow process and most formulations are unable to handle image rotation (but see [19]). In our approach we bypass the need for image deblurring by selecting a sharp video.

2.2. Stream Selection

Selecting the best surveillance camera out of a set of camera feeds has attracted much research interest. Most of the work is concerned with various criteria for determining the most interesting stream [17]. The static video camera setting is rather different from the egocentric video scenario and is therefore not directly applicable.

Some work [16, 15] has been done on combining videos (Video Mashup) from several mobile cameras. Their work has concentrated on learning video quality measures from professional Video Mashups. As the papers deal with handheld cameras, all streams are assumed to view the region of interest. This assumption does not hold for wearable egocentric cameras that are always on and mostly do not have viewfinders. Arev et. al. introduce in an upcoming paper [1] a method for producing a coherent video of an activity from multiple feeds of "social cameras". They use 3D reconstruction of the scene as well as the 3D poses of the cameras. This approach can generate good results, but 3D reconstruction may be very difficult in many cases. Unlike [1], our approach involves simple descriptors that do not need 3D reconstruction, and perform robustly in practice.

3. From Multiple Videos to a Single Video

Head mounted egocentric videos are often of poor quality due to large motions of human heads, causing videos to be jumpy and blurry. Video stabilization and image deblurring algorithms are usually unable to give satisfactory results for such challenging videos. Our objective in this work is to use several egocentric videos of the same event to create a single high-quality video. In Sec. 3.1 3.2 quality and popularity measures are defined for determining the desirability of each frame. In Sec. 3.3 dynamic smoothness costs are defined for transitions between video streams. The optimal set of frames is found efficiently by a dynamic programming method detailed in Sec. 3.4.

3.1. Video Quality Measure

Egocentric videos often contain periods of poor image quality. Large head motions are very common, resulting in

very unstable and sometimes blurry videos. For each frame in the video we calculate its SURF features [2] and track the feature points in the subsequent frame.

We define the shakiness of each frame t in stream v as $Q^v_{stab}(t)$, the average square displacement of all feature points between frame t and frame $t+1$:

$$Q^v_{stab}(t) = \sqrt{(dx^v_{mean}(t))^2 + (dy^v_{mean}(t))^2} \quad (1)$$

Frames with small movements are therefore strongly preferred to frames with large movements. It is possible to first stabilize each individual video separately but we have not found it necessary. The resulting video is good whether each input video is stabilized or not.

Another significant factor affecting image quality is motion blur. Under fast motions images can be significantly degraded. Following [9] we use the peakiness of the gradient distribution as our blur measure. Specifically we use the ratio between the 90th percentile and the 99th percentile. Lower scores indicate more peaked distributions for higher gradients and therefore sharper images.

$$Q^v_{sharp}(t) = \frac{percX^v_{90}(t)}{percX^v_{99}(t)} + \frac{percY^v_{90}(t)}{percY^v_{99}(t)} \quad (2)$$

Where $percX^v_m$ indicates the m th percentile of the absolute gradient in the x direction and the same follows for y. By choosing frames with low $Q_{sharp}(t)$ scores, sharp frames are preferred.

3.2. Popular is Interesting

Determining the interesting frames in a single video is a very challenging task requiring a good degree of scene understanding. Current methods fall far short of such capabilities. An important contribution of our paper is utilizing a collection of sequences captured by different users for accurately identifying interesting frames. When sitting in a lecture for example, users often stare at objects that are not of primary interest to other users such as their watches or at their books. Such frames will frequently be of high image sharpness and stability. For the single video case the importance of such frames is unclear whereas for the multiple users case they are likely to be considered as uninteresting.

Our approach is to use multiple cameras for determining the interestingness of each frame by evaluating its "popularity". Popularity is defined by the number of cameras looking at the same time at the same scene region for a long enough period of time.

To compute the popularity of frames at some time t, we must determine how many cameras are looking at the same region. As the baseline between different observers is often very wide, we have found that image comparison by keypoint matching is not always robust, instead we match frames based on their color histograms.

For each frame $F^v(t)$ in video v at time t we compute 256 bin histograms in each of their chromatic components (Cb and Cr). We denote the histogram operation as $H_{Cb}(), H_{Cr}()$ for the two components. We further define the Earth Movers Distance (EMD) between two frames as the sum of the EMD [18, 14], computed separately for each chromatic component. This is given in Eq. 3:

$$EMD(A,B) = EMD(H_{Cb}(A), H_{Cb}(B)) + EMD(H_{Cr}(A), H_{Cr}(B)) \quad (3)$$

We model each video stream v to consist of several sequences of frames, each containing observations of one of the K regions of interest in the scene ($R^v_k, k = 1:K$). For example in classrooms or concerts $K = 1$, as we have a single area of interest. In a meeting of 3 people $K = 2$ as each person can the see the other two participants. There are also numerous anomalies where the frames do not contain a region of interest (they might contain images of the window or other uninteresting regions). Mathematically this is formulated as a HMM, where the observable variable is the color histogram of a frame $F^v(t)$ and the hidden variable is the identity of the region of interest R^v_k viewed by the frame, $F^v_{Label}(t)$. The unnormalized log (emission) probability of a histogram given a region of interest is defined by:

$$log(P(H(F^v(t))|H(R^v_k))) = -min(\frac{EMD(H(F^v(t))|H(R^v_k))}{\sigma}, \tau) \quad (4)$$

Where the estimation procedure for the center histogram of the ROI $H(R^v_k)$ is described below. The mismatch cost between frame and cluster is capped at τ, modeling frames in which anomalous regions are viewed for a short period. σ is the $\frac{20}{K}\%$th percentile of the distances between cluster centers and frame histograms, describing the variability in distances around cluster centers.

Most adjacent frames in a video stream observe the same region of interest. To encode this in the HMM, we impose a Potts transition probability C_{Smooth} on switching region of interest labels between temporally adjacent frames. The unnormalized log probability of transition between labels $F^v_{Label}(t)$ and $F^v_{Label}(t+1)$ is defined as below:

$$log(P(F^v_{Label}(t+1)|F^v_{Label}(t))) = \begin{cases} 0 & \text{if} F^v_{Label}(t) = F^v_{Label}(t+1) \\ -C_{Smooth} & \text{otherwise} \end{cases} \quad (5)$$

As we do not know the ROI centers we need both to estimate the ROI center histograms $H(R^v_k)$ and infer the frame labels $F^v_{Label}(t)$. We do this using an EM type procedure. We initialize the ROI center histograms by clustering

the histograms of each stream v into K centers using K-Means. We use these values as initial guesses for the color histograms of the K scene regions of interest in each video v.

Finding the optimal MAP labeling of frames $F^v_{Label}(t)$ can be done efficiently using standard methods [3]. The MAP labeling assigns each frame t in stream v to one of the ROIs R^v_k. Frames with distance from their assigned cluster centers larger than τ are assigned as anomalous. Using the new cluster assignment we recompute the ROI centers, anomalous frames are not used for recalculating cluster centers. The above procedure (MAP inference and center re-computation) is repeated until convergence, in practice it only takes around 3 iterations. The final result is a set of ROI centre histograms for each stream $H(R^v_{1..K})$, and frame assignment to clusters $F^v_{Label}(t)$.

We proceed to find the identity of ROIs across different streams corresponding to the same scene region. As we do not know for certain which ROIs in different streams $R^{v_1}_i$,$R^{v_2}_j$ correspond to the same scene region, we propose a measure $d_{match}(R^{v_1}_i, R^{v_2}_j)$ for the likelihood of a match. The measure has chromatic and temporal overlap components. The chromatic component is given by the EMD between the ROI centers corresponding to similarity of color content (normalized by the distance between the most similar clusters). The temporal overlap component is given by the ratio of temporal overlap between the times of frame assigned to the ROIs and the minimal number of frames assigned to the ROI. We assume that frames assigned to ROIs in different cameras corresponding to the same scene region will be mostly overlapping in time. Formally this can be written as:

$$d_{match}(R^{v_1}_i, R^{v_2}_j) = EMD(R^{v_1}_i, R^{v_2}_j) + \lambda \cdot (1 - \frac{overlap(R^{v_1}_i, R^{v_2}_j)}{minDuration(R^{v_1}_i, R^{v_2}_j))} \quad (6)$$

Where λ is a parameter (we use 3). The affinity matrix A is defined by $A^{v_1,v_2}_{i,j} = e^{-d_{match}(R^{v_1}_i, R^{v_2}_j)}$.

Finding the most likely assignment of ROIs across different streams can be cast as a clustering problem. We solve it using Normalized Cuts with affinity matrix A and K clusters. The optimal assignment induces a set of scene labels S_k,$k = 1..K$ which can be assigned to each R^v_k. Note that scene labels are independent of the stream.

Finally we can denote the popularity score for frame $F^v(t)$ as the number of frames at time t sharing the same label S_i.

$$C^v_{pop}(t) = \#\{\tilde{v} | F^{\tilde{v}}_{Label}(t) = S_i\} \quad (7)$$

3.3. Smoothness Measure for Transition Cost

Although choosing the frame of the best quality and popularity at each time interval results in the best possible score, it is obvious that switching very frequently will result in a jumpy video with frequent abrupt transitions. As we would like to discourage overly frequent transitions, we associate a certain cost with each transition between streams (i.e. when frame $t+1$ comes from a stream different from that of frame t). We denote the cost of transitions between streams v_1 and v_2 at time t, $C^{v_1,v_2}_{Smooth}(t)$.

Due to spatial proximity, transitions between more closely overlapping streams are preferred to more detached streams e.g. streams both showing the lecturer‘s head rather than one showing his head and the other showing his feet. Overlap is computed by matching feature descriptors between frames (we use SURF [2]) and estimating an affine transformation between them. Although not as robust as color histograms (thus not used as a popularity measure), we have elected to use keypoint matching as it is a better measure of overlap. Due to geometric and radiometric differences between streams, not all frames can be reliably matched. For such cases we default to a constant cost equivalent to an overlap of 0. The overlap is computed as in Eq.8:

$$O^{v_1,v_2}(t) = \frac{F^{v_1}(t) \cap F^{v_2}(t)}{F^{v_1}(t) \cup F^{v_2}(t)} \quad (8)$$

Where $F^{v_1}(t) \cap F^{v_2}(t)$ denotes the intersection between the areas of frame t of videos v_1,v_2 and $F^{v_1}(t) \cup F^{v_2}(t)$ corresponds to their conjunction.

We therefore choose the cost of transition between two streams at time t to be:

$$C^{v_1,v_2}_{Smooth}(t) = C_s \cdot (2 - O^{v_1,v_2}(t)) \quad (9)$$

We use $C_s = 100$.

3.4. Optimization

In Sec. 3.1 3.2 we have detailed the suggested cost functions for selecting each frame. The cost is dependent on the quality and popularity of the frame. In sec. 3.3 we have detailed our smoothness cost function minimizing sharp transitions between video streams based on the overlap between frames.

The above cost function can be cast as a dynamic programming problem. Let $Z(t)$ denote the stream selected as time t. We define the singleton cost term $C_{Prior}(Z(t))$ as in Eq. 10:

$$C_{Prior}(Z(t)) = Q^{Z(t)}_{stab}(t) + \alpha \cdot Q^{Z(t)}_{sharp}(t) - \beta \cdot C^{Z(t)}_{pop}(t) \quad (10)$$

The pairwise cost between subsequent $Z(t)$ values is given by the smoothness cost as in Eq. 11

$$C_{Pair}(Z(t), Z(t+1)) = C^{Z(t),Z(t+1)}_{Smooth}(t) \quad (11)$$

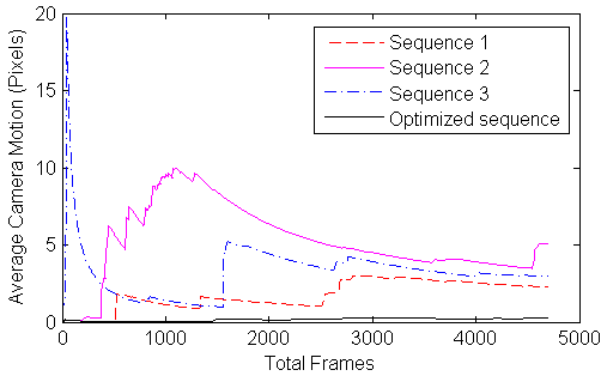


Figure 2: The effect of our method on the stability of the sequences. The RMS camera motion of the 3 sequences is 3.44 pixels per frame whereas the RMS camera motion in our optimized sequence is only 0.25 pixels per frame. This is an improvement of 13X without the usage of any stabilization algorithm. The x-axis represents the frame number from the start of the sequence and the y-axis represents the RMS camera motion, measured by pixels per frame.

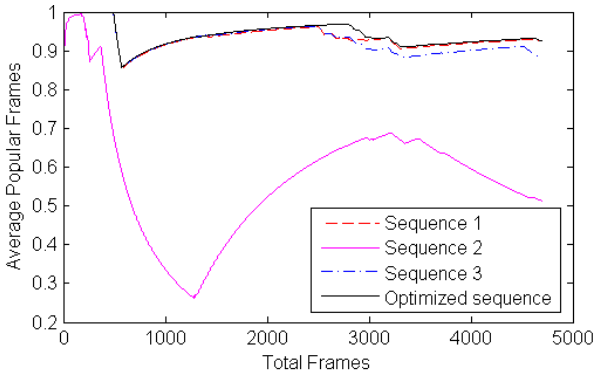


Figure 3: The effect of our method on the popularity of the sequences. The average fraction of important frames viewed by the 3 sequences is 0.77 whereas the average fraction of important frames viewed in the output sequence is 0.93. The x-axis represents the number of frames from the beginning of the sequence and the y-axis represents the average fraction of important frames included in the sequence.

The total cost is the sum of all singleton and pairwise costs as given by Eq. 12:

$$C_{Total}(Z(1)..Z(T)) = \sum_t C_{Prior}(Z(t)) + C_{Pair}(Z(t), Z(t+1)) \quad (12)$$

The cost function can be optimized exactly by dynamic programing [3] with complexity $O(TV^2)$, where T is the number of frames in each stream and V the number of streams.

The output video is the temporally ordered sequence of frames, where the frame at time t is selected from stream $Z(t)$.

4. Experiments

We present several experiments demonstrating the performance of our method in relevant scenarios. All scenario were taken by 3 synchronized head worn egocentric cameras (GoPro HERO 3 and 3+) The different scenarios were recorded by different participants. After recording the scenarios, the videos were processed by our method detailed in Sec. 3. Smoother transitions were obtained by implementing a linear-fade effect at the transition between camera streams in the output video.

The scenarios are detailed below:

- Scene A - Lecture - An undergraduate lecture was recorded by 3 cameras. One of the users alternates between writing notes and looking at the lecturer. Each of the videos experiences sharp transitions and periods of shakiness. The output video is stable, contains footage from all 3 cameras and only shows frames of interest (those containing the lecturer)

- Scene B - Concert - An open air musical concert was recorded by 3 cameras. Some of the users look at areas other than the performer and all experience periods of severe shakiness. The output video is of much higher quality and only shows frames containing the performer. Frames from only two cameras were selected as one of the users is a much shakier photographer than the others.

- Scene C - Seminar - A seminar in a small classroom was recorded by 3 cameras sitting in the second row of chairs. The lighting is much darker than in the other scenarios. The output video is of higher quality than the component videos and only displays the region of interest (the speaker).

  The complete videos can be seen on `https://www.youtube.com/channel/UCce6USxoqtBh9-MRRnnlQwg/videos`

  In Fig. 2 the stability score of each input video frame is plotted as a function of time for all streams. The stability of the output video generated by our method is also shown on the graph. The root mean squares (RMS) movement between adjacent frames is used to measure the stability of the sequence. The average RMS stability of the input videos is 3.44 pixels, whereas the RMS of output video is 0.25 pixels a 13X improvement. It is readily seen from the graph that the output video is better than all input videos.

  In Fig. 3 the popularity score of each of the input and

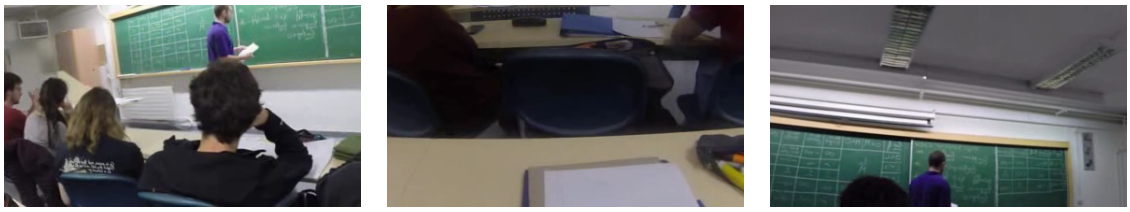
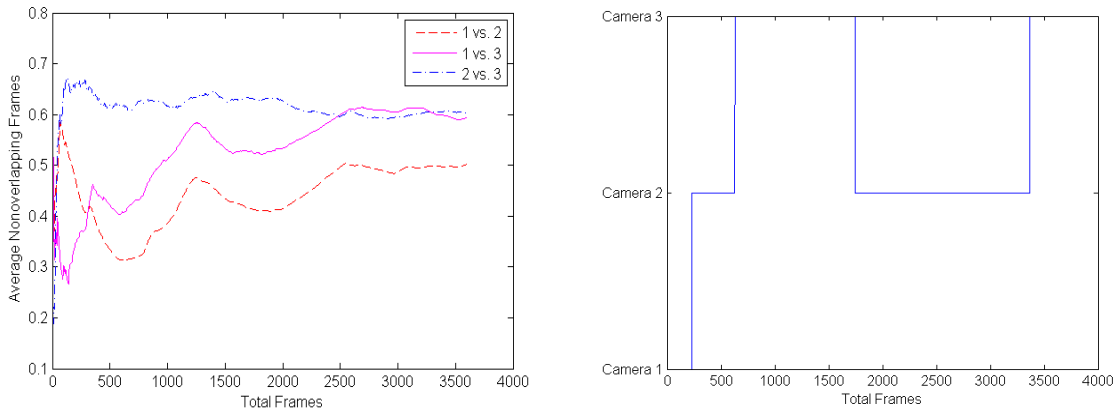


Figure 1: Example result of the Popularity Measure. The left and the right frames were marked by our algorithm as 'popular' whereas the middle frame was marked as 'unpopular'. The popularity measure quantifies the importance of a frame. Unpopular frames are not included in output sequence and therefore the identification of popular frames is an essential part of our optimization procedure.



a. b.

Figure 4: The Smoothness Measure. (a) The measure of similarity between the frames at time t for each of the cameras based upon their overlapped field of view (see text). The similarity is measured between 0 to 1, where 1 indicates identical frames. Sequence 2 and sequence 3 are very similar. (b) The transition cost between sequence 2 and 3 is less than the transition cost to sequence 1 and we can see that the optimized sequence indeed exhibits more transitions between these sequences.

output video frames is plotted as a function of time. Low popularity frames are very rarely selected in the output video. The average popularity score within the input videos is 0.77 as opposed to 0.93 in the output video. The output video frames are more popular than all component videos. This demonstrates that in representative scenarios, our algorithm is able to choose frames that are both popular and of high quality.

5. Conclusion

We have presented a method for creating a single high quality video of an event from several variable quality wearable egocentric videos. This was done by selecting at each time instant the highest quality frame which the crowd deemed interesting. Most current methods are suitable for mobile phones and consider all frames to be interesting. As egocentric cameras are always on we have no guarantee that all frames are interesting. In fact many parts of each input video contain unwanted footage such as staring at the window or glancing at the user's mobile phone. By detecting the parts of the scene that most users looked at at any given time we were able to find the frames of interest.

We suggested measures for both quality and popularity of each frame, and an overlap based technique for smoother transitions between shots. The frame selection problem was formulated as an energy minimization problem and efficiently optimized using dynamic programming. Real-life experiments were presented, demonstrating the effectiveness of the method in several scenarios of interest.

We believe our method will be highly beneficial for creating high-quality footage in scenes that at the present time are recorded at sub-par quality or not at all e.g. small lectures, seminars, amateur concerts and events.

At the moment our method uses no priors specific to the event recorded e.g.frames containing no moving

a. b. c.

Figure 5: The sequences we have used in the experiments.(a) Lecture, (b) Concert, (c) Seminar.

objects are usually of low interest, prefer frames with the object of interest in the center. It is very likely that learning such priors from data would improve frame interestingness further. This is left as a future research direction.

Acknowledgement: This research was supported by Google and by the Israel Ministry of Science.

References

[1] I. Arev, H.-S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, page to appear, 2014.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417. 2006.

[3] R. Bellman. Dynamic programming and lagrange multipliers. *PNAS*, 42(10):767, 1956.

[4] C. Buehler, M. Bosse, and L. McMillan. Non-metric image-based rendering for video stabilization. In *CVPR*, volume 2, pages II–609, 2001.

[5] S. Cho and S. Lee. Fast motion deblurring. In *ACM TOG*, volume 28, page 145, 2009.

[6] A. Goldstein and R. Fattal. Video stabilization using epipolar geometry. *ACM TOG*, 31(5):126, 2012.

[7] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *Proc. of the 21st ACM int. conf. on Multimedia*, pages 1017–1026, 2013.

[8] M. Grundmann, V. Kwatra, and I. Essa. Auto-directed video stabilization with robust l1 optimal camera paths. In *CVPR*, 2011.

[9] D. Krishnan, T. Tay, and R. Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, pages 233–240, 2011.

[10] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, pages 1964–1971, 2009.

[11] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, pages 3570–3577, 2013.

[12] F. Liu, M. Gleicher, J. Wang, H. Jin, and A. Agarwala. Subspace video stabilization. *ACM TOG*, 30(1):4, 2011.

[13] Y. Matsushita, E. Ofek, X. Tang, and H.-Y. Shum. Full-frame video stabilization. In *CVPR*, volume 1, pages 50–57, 2005.

[14] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *ICCV*, 2000.

[15] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. Movimash: online mobile video mashup. In *Proc. of the 20th ACM int. conf. on Multimedia*, pages 139–148, 2012.

[16] P. Shrestha, H. Weda, M. Barbieri, E. H. Aarts, et al. Automatic mashup generation from multiple-camera concert recordings. In *Proc. of the 18th ACM int. conf. on Multimedia*, pages 541–550, 2010.

[17] S. Sumec. Multi camera automatic video editing. In *Computer Vision and Graphics*, pages 935–945. Springer, 2006.

[18] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multidimensional histograms. *CVGIP*, 1985.

[19] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. *IJCV*, 98(2):168–186, 2012.