International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Viewpoint Selection for Human Actions

Dmitry Rudoy · Lihi Zelnik-Manor

Received: date / Accepted: date

Abstract In many scenarios a dynamic scene is filmed by multiple video cameras located at different viewing positions. Visualizing such multi-view data on a single display raises an immediate question - which cameras capture better views of the scene? Typically, (e.g. in TV broadcasts) a human producer manually selects the best view. In this paper we wish to automate this process by evaluating the quality of a view, captured by every single camera. We regard human actions as three-dimensional shapes induced by their silhouettes in the space-time volume. The quality of a view is then evaluated based on features of the space-time shape, which correspond with limb visibility. Resting on these features, two view quality approaches are proposed. One is generic while the other can be trained to fit any preferred action recognition method. Our experiments show that the proposed view selection provide intuitive results which match common conventions. We further show that it improves action recognition results.

Keywords Video analysis · Viewpoint selection · Human actions · Multiple viewpoints

1 Introduction

With the advances of recent years video cameras can now be found in abundance. Scenes and events are frequently being recorded not only by a single camera, but rather by multiple ones, e.g., school children sports events are recorded by many eager parents and street shows are filmed by enthusiastic tourists. There are several possible ways to visualize such data. Ballan et al. in (2010) proposed a 3D reconstruction based visualization system that allows smooth switching between views. To visualize such data on a single screen one needs to select the single "best" camera for each moment in the event. In movie and TV production the "best" camera view is selected manually by a human producer. Such a producer is typically not available in non-professional scenarios. Therefore, we are interested in automating the process of camera selection.

In this paper we propose a technique for video-based evaluation of the quality of a view for actions. We first discuss what makes one view better than the other. Our guiding principle is that the better views are those where the action is easier to recognize. We then present the properties of such views, propose three measures (spatial, temporal, and spatio-temporal), which capture them and incorporate them into a single global score. Since our goal is to detect views where the action is recognizable, we further propose an approach for learning to detect the good views for a particular recognition method.

The usefulness of our view selection is evaluated qualitatively on real video data of sports events, dance and basic human actions. Additionally we test our approach on 3D gaming scenarios. To provide some quantitative evaluation we further test the usefulness of the proposed approach for action recognition. Here, rather than using all views, we use only the better ones for recognition. Our experiments show that selecting a single good view to process does not deteriorate recognition rates, but rather the opposite occurs and recognition rates are improved. This could speed-up recognition in multi-camera setups.

D. Rudoy
EE dept., Technion, Israel
Tel.: +972-4-8294680
Fax: +972-4-8295757
E-mail: dmitryr@tx.technion.ac.il

L. Zelnik-Manor
EE dept., Technion, Israel

The contribution of the paper is hence threefold. First, it presents several properties of preferable views of human actions (Section 3). Second, two methods are proposed for capturing these properties and hence estimating the relative quality of the views. In Section 4 we present a generic view selection approach, while in Section 5 we train the view selection to match a given recognizer. Last, we demonstrate the benefits of the suggested approach in scene visualization and establish the selection of the better views by testing on action recognition (Section 6).

Fig. 1 Many actions are better captured from a specific view point. Walking and hugging are best captured from the side, while a golf swing is best viewed from the front. Top row: examples of road signs. Bottom rows: YouTube search results for "hugging people" and "golf swing".

2 Related work

Viewpoint selection was previously addressed in many different fields. An extensive review of view selection in computer graphics is available in (Christie et al., 2008). Several methods, e.g., by Mudge et al. (2005), Vazquez et al. (2003), Bordoloi and Shen (2005) have been proposed for optimal view selection for static 3D objects. Vieira et al. (2009) utilize a learning scheme for interactive view selection. Assa et al. in (2008; 2010) proposed methods for camera control when viewing human actions. These solutions, however, rely on knowing the 3D structure of the scene and, hence, are not applicable to real-world setups filmed by video cameras.

Camera selection has also been previously explored for surveillance applications, however, there the camera setup and the goal are typically different from ours. Gorshorn et al. (2007) propose a camera selection method for detecting and tracking people using camera clusters. Most of the cameras in the cluster do not overlap and hence the main goal is tracking the person as he moves from one camera view to the other. In (Krahnstoever et al., 2008; El-Alfy et al., 2009) methods were proposed for selecting the camera that provides the best view for recognizing the identity of a viewed person. This requires mostly face visibility.

An example of multiple view scene visualization system was proposed by Ballan et al. in (2010). They reconstruct a model of the background, estimate camera positions for every frame and then place each video stream at it's exact 3D location. This allows the user to view all the videos simultaneously on the 3D model of the environment, or switch between videos according to the cameras' locations. The viewpoint quality measures suggested below could be used to automate viewpoint transitions and hence improve the user's experience.

3 Why some views are better than others?

Many human actions are easier to recognize from some viewpoints, compared to others, as illustrated in Figure 1. This is why "WALK" road-signs always show the stride from the side, YouTube "golf swings" are almost all frontal views while "hugging people" videos are mostly side views showing both people approaching each other. The mutual to all these examples is that the better views are those showing the limbs and their motion clearly. Hence, these visibility criterions have been used for camera path planing in computer animation (Assa et al., 2008).

Based on this observation, our goal is to evaluate limb visibility. We wish to achieve that without detecting or tracking the limbs, since limb detection is time consuming and error prone. Instead, we observe that good visibility of the limbs and their motion has generic temporal, spatial and spatio-temporal implications on the space-time shape induced by the silhouettes (as illustrated in Figure 2):

1. Temporal: High motion of the limbs implies that the silhouettes vary significantly over time.
2. Spatial: Good visibility of the limbs implies that the outlines of the silhouettes are highly concave. For example, the spread out legs in a side view of walk generate a large concavity between them.
3. Spatio-temporal: When the limbs and their motion are clearly visible the resulting space-time shape is not smooth, but rather has protruding salient parts (corresponding to the moving limbs). Conversely, self occlusions and lack of motion lead to smooth space-time shapes.

Interestingly, each of these three properties matches known properties of human perception. First, it is known that human vision is attracted to high motion (Johansson, 1973). This corresponds to the temporal property.

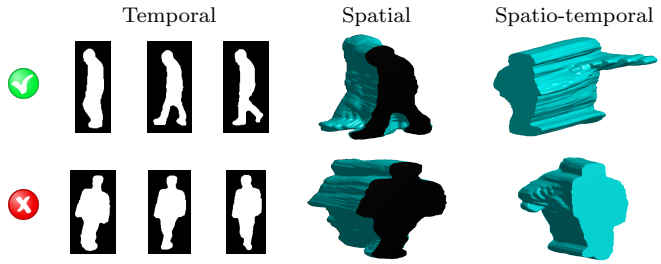


Fig. 2 Properties of good views. Good views are those where the limbs and their motion are clearly visible. Temporally this implies large variations. Spatially, the silhouettes in a good view are highly concave. In space-time visibility of the limbs and their motion implies shapes with significant saliency.

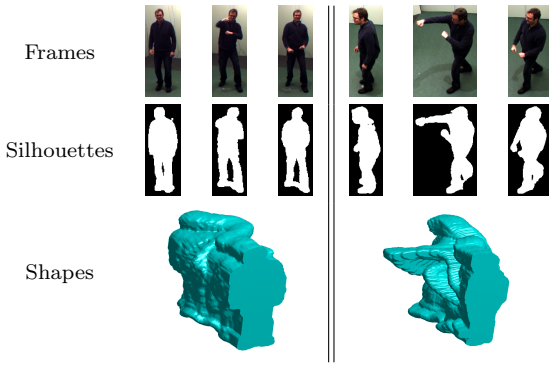


Fig. 3 Example of space-time shapes of the same action from different viewpoints. Given video frames (top row) the human figure is extracted (middle row). By aligning the center of mass of the silhouettes the space-time shape (bottom row) is created. Note the differences between the shapes obtained for different viewing positions.

Second, Attneave (1954) proposed that the most informative regions along a visual contour are those with high curvature. Extending this idea to three dimensions matches the spatio-temporal property, that looks for protruding regions with high curvature in space-time. Finally, Feldman and Singh showed in (2005) that for closed planar contours concave segments carry greater information, than corresponding convex ones. Correspondingly, the presented spatial property captures the concavity of the silhouettes.

For the time being we limit our analysis to scenes showing a single person performing free actions. Later on, in Section 6.1 we discuss how multi-player scenes are handled.

4 Measures of viewpoint quality

In this section we propose measures for evaluating the quality of a viewpoint, which are based on the principles presented above. We intentionally seek for simple measures of viewpoint quality to enable fast processing.

First, we assume that each video can be segmented into foreground and background using background subtraction and that the cameras are fixed. Then, following (Gorelick et al., 2007) we compensate for the global translation of the body in order to emphasize motion of parts relative to the torso. This is done by aligning the centers of mass of the silhouettes to the same point. This may introduce non-existent motions, but we found out that they are less important than the global one. The process is illustrated in Figure 3 and results in a three-dimensional shape induced by the actor's silhouettes in the space-time volume. Note, that cameras at different positions view different silhouettes, hence, the induced space-time shapes are different. We assume that all the cameras view the person fully without external occlusions. Self occlusions (e.g., as in a top view) are allowed.

Our measures do not require perfect silhouettes, however, we do assume the silhouettes are acceptable, i.e., when there are no self occlusions the limbs should be visible in the silhouette. This assumption is reasonable in many scenarios, e.g., computer games, day-time sports events and security setups where the background can be modeled accurately and does not change much. We will address this issue later, in Section 7.

4.1 Spatio-temporal measure: Shape saliency

In accordance with property (3), when the limbs are visible the induced space-time shape is not smooth. To build a spatio-temporal measure of action visibility we need to quantify the unsmoothness of the space-time shapes. We base our approach on the method proposed by Lee et al. (2005) for evaluation of saliency and viewpoint selection for 3D meshes. Their work proposes a method for measuring saliency at every vertex of a static 3D mesh. Saliency is defined as the deviation of the mesh from a perfectly smooth shape, i.e., sphere. Furthermore, they propose to evaluate a viewpoint quality by summing up all the saliency values of all the visible parts from a given viewpoint. In our work instead of the same shape viewed from different directions we have a different shape for each view. Thus measuring the overall saliency of the space-time shape allows us to estimate the quality of the view that produced that shape.

Following (Lee et al., 2005) we first calculate the local space-time saliency at each point on the shape's surface. This is captured by the difference between the point's local curvature at different scales. Then we evaluate global saliency by summing all the local saliency values, since every point on the surface of the space-

time shape is visible. The method in (Lee et al., 2005) was limited to 3D meshes. In our case, however, the shapes are represented in voxels and not meshes. We next follow the ideas of (Lee et al., 2005) and extend them to voxel-base representations of space-time shapes.

To compute the local saliency of points on the surface of the space-time shape we first calculate the mean curvature $\kappa_m(p)$ of each surface point p using the method proposed by Kindlmann et al. (2003). In a nutshell, their method uses convolution with a continious filter for curvature computation instead of parametrisation of the surface. Further details of the method can be found in (Kindlmann et al., 2003). Next, following Lee et al. (2005) we define the weighted mean curvature, $G(\kappa_m(p), \sigma)$, at each space-time shape surface point, as:

$$G(\kappa_m(p), \sigma) = \frac{\sum_q \kappa_m(q) W(p, q, \sigma)}{\sum_q W(p, q, \sigma)}, \quad (1)$$

where the sum is over all the points q within a 2σ radius neighborhood around point p and $W(p, q, \sigma)$ is a weight function. Note, that as opposed to the 3D models used in computer graphics, our shapes can have different scales in space and in time. Hence, we define $W(p, q, \sigma)$ as:

$$W(p, q, \sigma) = exp\left(-\frac{1}{2}\left(\sum_{j \in \{x,y,t\}} \frac{(p_j - q_j)^2}{\sigma_j^2}\right)\right), \quad (2)$$

where $p = (p_x, p_y, p_t)$ and $q = (q_x, q_y, q_t)$ are two points on the space-time shape surface, and $\sigma = (\sigma_x, \sigma_y, \sigma_t)$. Local space-time saliency is then defined as the absolute difference between two weighted curvatures:

$$L(p) = |G(\kappa_m(p), \sigma) - G(\kappa_m(p), 2\sigma)|. \quad (3)$$

For more details the reader is referred to (Lee et al., 2005).

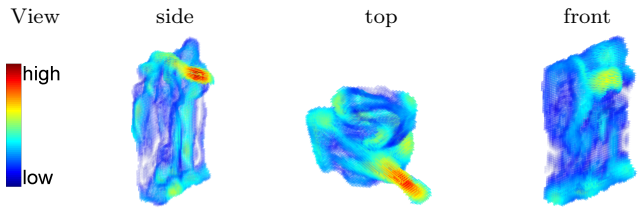
Lee et al. (2005) propose to incorporate multiple scales, but in our space-time shapes this doesn't bring the desired effect. This is because 3D models usually have many small details, that need to be taken into account. In contrast, our space-time shapes are very low detailed, thus, it would suffice to find a single optimal value of σ corresponding to a single scale. We have experimentally selected $\sigma = (5, 5, 3)$. This value was kept fixed for all the results presented here.

Figure 4 shows the local space-time saliency values of Equation (3) for all the surface points of the space-time shape of a real punch action obtained from different views. It can be seen that the moving parts, e.g., the arm, generate high curvature surfaces and hence receive high local saliency values, while stationary parts, e.g., the legs, which imply flat areas in the space-time shape, receive low local values.

Finally, we define the spatio-temporal visibility score of a view as the sum of the local saliency values of all the points on the surface of the space-time shape:

$$S_{ST} = \sum_p L(p), \quad (4)$$

The values of the spatio-temporal saliency score S_{ST} for a punch action are also marked in Figure 4. We note here that S_{ST} is not bound from above, however it's always non-negative. In our C implementation computing S_{ST} takes 3 seconds on average for a sequence of 36 frames.



S_{ST} =	98.3	91.9	48.9

Fig. 4 Spatio-temporal visibility measure. Local saliency values for space-time shapes obtained from different views of the same punch action nicely emphasize the more important regions, in this case, the punching arm. The side and top views show the protruding punching arm and hence their total visibility score S_{ST} is high. The front view produces a smooth shape and correspondingly a low S_{ST}.

As noted above the S_{ST} score is not normalized and can receive any non-negative values. We have tried several different normalizations, however, our experiments showed that the un-normalized measure performes better than the normalized ones. Additionally, we have tested with several other methods to capture saliency, e.g, the one proposed by Gorelick et al. (2007), but all performed worse.

4.2 Spatial measure: Visibility of limbs

According to property (2), when the limbs are fully visible the outlines of the induced silhouettes are highly concave. To quantify how concave a shape is we seek for a simple and fast measure. One such measure is computing the volume difference between the 3D convex hull of the space-time shape and the shape itself. We define the spatial measure as:

$$S_S = 1 - \frac{V_{sh}}{V_{ch}}, \quad (5)$$

where V_{sh} is the volume of the space-time shape and V_{ch} is the volume of its convex hull. The S_S score is

non-negative and bounded by 1 from above. Figure 5 illustrates some space-time shapes together with the corresponding 3D convex hulls. Note, that the computation of this score is fast and takes 0.3 seconds for a 36 frame long clip.

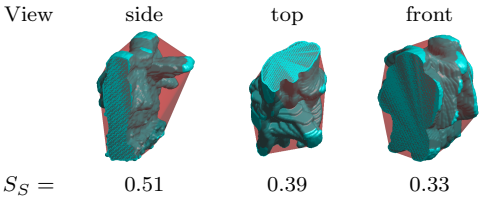


Fig. 5 Spatial visibility measure. An illustration of a space-time shape (cyan) and its convex hull (red) for a "punch" action captured from different angles. The side view receives the highest S_S score since it shows more concave regions under the arm. The top and front views, where the limbs are less visible, obtain lower scores, since their outlines are more convex.

4.3 Temporal measure: Detecting large variations

Following property (1), we wish to discover views which exhibit a significant pattern of change along time. We measure this by computing the portion of pixels where motion was observed somewhere along the sequence. Note, that we do not care what was the type of motion or when it occurred. Our only interest is the amount of motion. Since we would like our measures to be as simple and fast to compute as possible, we evaluate the amount of motion as follows. Let $g(x, y, t)$ be the silhouette indicator function and τ be the temporal length of the sequence. Following (Bobick and Davis, 2001) we build the motion-energy image describing the sequence as:

$$E(x, y) = \bigcup_{t=0}^{\tau-1} g(x, y, t). \quad (6)$$

We denote by $g_m(x, y)$ the biggest single-frame silhouette (in sense of number of pixels) in the sequence. The temporal measure is then defined as:

$$S_T = 1 - \frac{\sum_{x,y} g_m(x, y)}{\sum_{x,y} E(x, y)}. \quad (7)$$

Note, that this score is always non-negative and bounded by 1 from above. Computing S_T takes approximately 2 miliseconds for a video of length 36 frames.

To illustrate the temporal motion-based score we present in Figure 6 $E(x, y)$, $g_m(x, y)$ and S_T for different views of a kick action. As can be seen, the side view, where the action is better viewed, presents a higher percentage of moving pixels (gray), and thus receives a higher score.

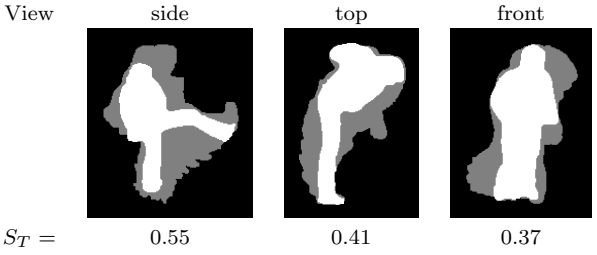


Fig. 6 Temporal visibility measure. The motion-energy image E (gray) superimposed by the biggest silhouette g_m (white) for a "kick" action seen from different views. The side view captures the leg motion in full and hence receives the highest score, while the front view shows very little of the leg motion and hence receives a low score.

4.4 Differences between measures

In the previous sections we presented three different visibility measures. Before using them in our experiments we wish to compare them. To do so we apply them to a variety of actions filmed from several angles. On one hand, the measures capture similar properties of action visibility, thus frequently support each other, as shown in Figure 7. However, even in those cases the measures emphasize differently the quality of the views. On the other hand, there areactions for which the measures are not consistent, see Figure 8. This occurs mainly when there are several views that provide good visibility of the action. In those cases using all measures for the view selection could yield more robust results.

4.5 Combining all measures

The above presented measures capture somewhat different notions. To take advantage of them all we further combine them into a single score. To accomplish this without parameter tuning we take the product of the three, yielding:

$$S_c = S_{ST} \cdot S_S \cdot S_T. \quad (8)$$

In our experiments we tried other combinations that do require parameter, however, no significant improvement was achieved.

By now we have proposed three different visibility measures and their combination. To illustrate their performance in a simple case we use the "Kung-Fu Girl"

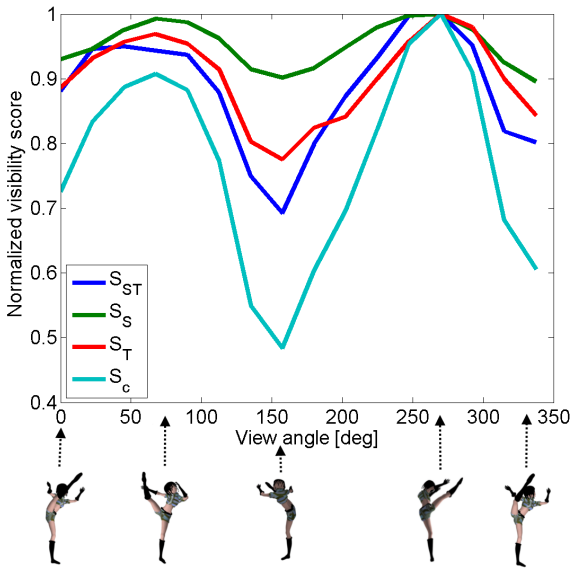


Fig. 7 Comparison of the different visibility measures for a kick action viewed by 16 cameras. In this case all the measures rate the views consistently. It can be seen that all the measures get a clear maxima in the two side views, i.e., when the limbs are visible.

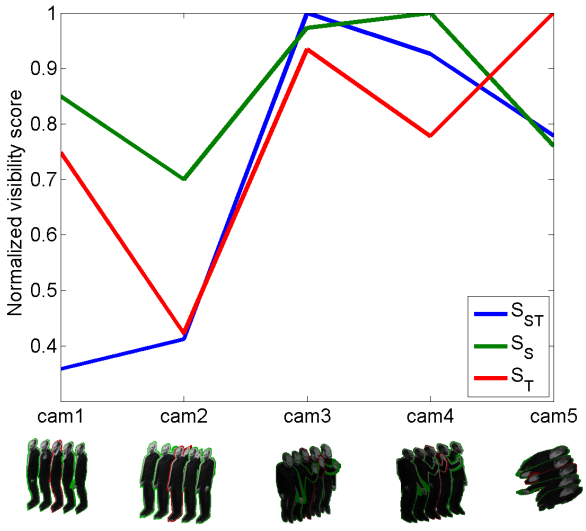


Fig. 8 Comparison of the visibility measures for a wave action captured by 5 different cameras. For this action the rating of the views are different for the different visibility measures.

dataset. [1] This dataset includes 25 videos of an animated 3D model of a girl performing a kung-fu drill. The viewpoints are distributed evenly on a hemisphere above the actress. Figure 7 shows how the viewpoint quality of the different measures changes as the viewing angle varies. It can be seen that there are clear maxima points in the side views for this specific kick. However, the combined measure S_c differentiates better between the best and the worst views (the gap between minima and maxima is larger), thus we expect it to provide a less noisy estimation on real data.

Figure 9 illustrates the result of applying this measure to the same "Kung-Fu Girl" dataset, but using all the cameras. According to one's intuition, side views, where the kick is best visible, receive high scores, while front and back views, where the arms and legs are occluded, receive lower scores.

5 Learning viewpoint quality

In the previous section we presented intuitive measures for capturing the properties of good views. While such model-based approaches often perform well, we further wish to examine whether one can learn to detect good views. An approach for doing that is described next.

Recall that our definition of a "good view" is one where the action is recognizable. Hence, to obtain labeled data of "good" (+1) and "bad" (−1) views we simply apply a recognizer to all videos in our dataset, in a leave-one-out manner. All videos for which recognition succeeded are labeled as "good" and those where recognition failed are marked as "bad". We then train an SVM-based quality estimator (Schölkopf and Smola, 2002) by representing each video with a vector $\mathbf{v}_i$:

$$\mathbf{v}_i = \left(\frac{S_T}{m_T}, \frac{S_S}{m_S}, \frac{S_{ST}}{m_{ST}}, \frac{S_c}{m_c} \right), \tag{9}$$

where $m_j = \max S_j$ for every $j \in \{S, T, ST, c\}$. The maximum is taken over the different views. The SVM learns the separation between the "good" and "bad" views in the kernel space (we use a Gaussian kernel with $\sigma = 2$) by presenting a function

$$\mathbf{v} \rightarrow \hat{s}. \tag{10}$$

$\hat{s}$ is positive for the "good" views and negative for the "bad" ones. The higher $\hat{s}$ is the better is the view. It is important to note that the quality measure of Equation (8) is generic and independent of the action recognition method to be applied afterwards. Conversely, the trained quality evaluation of Equation (10) depends on the initial labeling of the selected recognizer.

As noted above (Section 4.4) the measures are consistent in many cases, but not in others. Hence, we use all the measures together with their combination in the SVM training. Based on our experiments, this results in a view quality evaluator that is more rubust to the failures of any single measure.

[1] "Kung-Fu Girl" sequence is available at `http://www.mpi-inf.mpg.de/departments/irg3/kungfu/`.

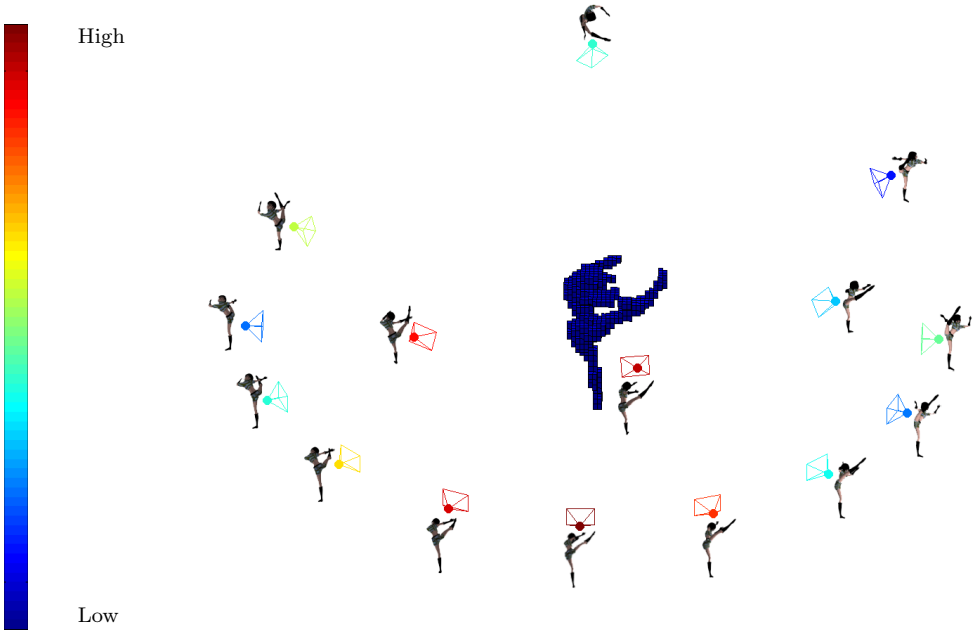


Fig. 9 Combined visibility measure calculated for every view of the "Kung-Fu Girl" sequence. To calculate the visibility measure we used a short (24 frames) sub-sequence around the presented frame. It can be seen that side views, where the arms and legs are visible, receive higher scores than front and back ones with self-occlusions. For clear exposition only half of available views are shown to prevent a mess.

6 Applications and Experiments

In this section we demonstrate the usefulness of the proposed view quality evaluation approach for simplified visualization of multi-camera scenes (Section 6.1). We cannot compare our view selection method to previous works, like (Assa et al., 2008, 2010), since they use 3D scene data, which is not available in our case, hence the evaluation is mostly qualitative. To provide some quantitative evaluation, we further show that selection of a single camera can improve action recognition rates (Section 6.2).

6.1 Automatic camera selection

The most intuitive application of view quality is automatic selection of the best camera. Given multiple video streams of the same scene filmed from different points of view, our goal is to produce a single video showing each action from its preferred viewpoint.

The quality of the view provided by a certain camera can change with time, depending on the performed actions and the person's orientation in space. To take those changes into account we adopt a sliding-window approach where view selection is applied to all sets of corresponding sub-sequences of length 36 frames. This yields 36 independent "best-view" decisions for each set of corresponding frames. We choose to work with relatively long sub-sequences since we wish to select the best viewpoint for the whole action and not for its fragments. For each frame the selected view is the one receiving the highest score S_c (of Equation (8)). To avoid redundant view switches we accept a view change only when it lasts longer than 25 frames.

Data captured in the lab. We begin by testing the proposed framework on a golf scene. We have intentionally selected golf since googling for "golf swing" videos retrieves many tutorials, most of them show the swing from the same frontal viewpoint, thus making it somewhat clear what is the desired result. In our setup eight cameras viewed a golfer hitting the ball four times, each time rotating to face a different camera. As shown in Figure 10 (a) and in the supplemental video our view selection approach successfully selects the frontal view for the swing, in line with what is used for golf tutorials.

Next we test the framework on a simple dance move. This move is best viewed from front, but the actress repeats it several times, each time facing a different direction. As in the golf scene, our view selection approach clearly prefers the front view. Other views are ranked according to the visibility of the limbs motion, as shown in Figure 10 (b). Note that the differences between the score values on the golf swing and the dance move orig-

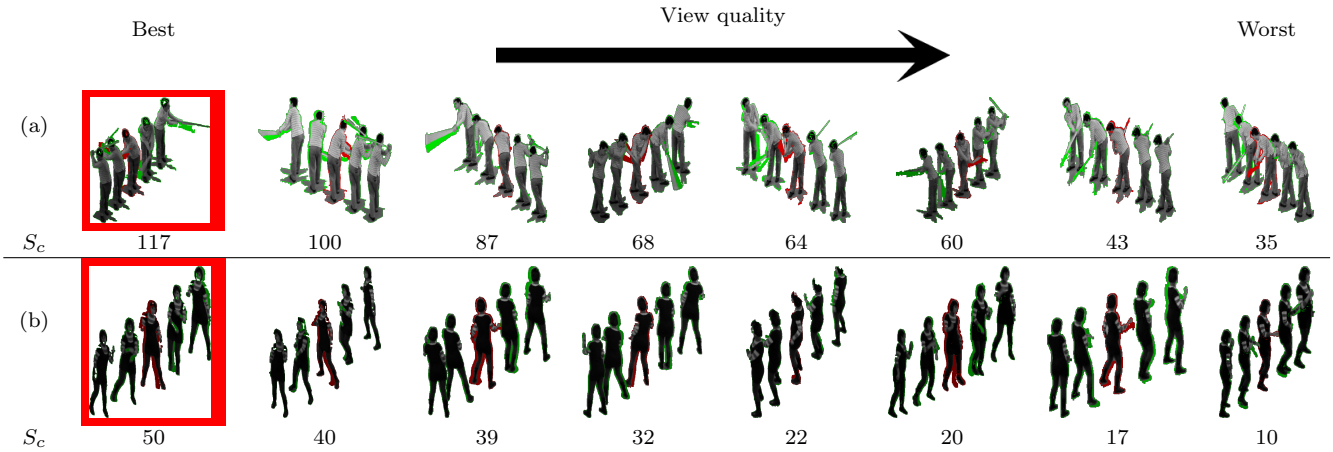


Fig. 10 View selection for golf swing and dance. Views showing clearly all the limbs together with their motion are ranked higher while views with severe self-occlusion are detected as low quality.

inate from the differences in the nature of the actions. The swing is a faster action thus it yields a higher score.

We further applied the proposed view selection to the IXMAS dataset (`IXMAS`), which includes 12 actors performing 13 everyday actions continuously. Each actor performs the set of actions three times, and the whole scene is captured by 5 synchronized video cameras (4 side cameras, that cover almost half a circle around the subject and one top camera). The actors selected freely their orientation, hence, although the cameras were fixed, each viewed the actors from varying angles. In other words, we cannot label a certain camera as front view since it captured both front and side views. In this experiment we don't use the ground-truth provided with the database.

As illustrated in Figure 11 and in the supplemental video, our system consistently selects views where the action performed by the person is clearly visible. For example, for walking the algorithm selects the side view with the maximum visibility of the moving legs, and for waving the front view is selected, such that the hand motion is clearly visible. Note, that since people oriented themselves freely, different cameras are selected for different people. As mentioned previously, the score differences originate in the activity differences. Additionally, here the human figure is considerably smaller than in the golf and dance scenes thus the scores are lower.

Obviously the proposed view selection method is not perfect. Some of the failure cases are shown in the Figure 12. In some cases the action is not suitable for view selection based on silhouettes only thus our method prefers incorrect views (Figure 12 (a)). In other cases there are many other motions, like arms motion in Figure 12 (b) along with the main action. Those additional motions can dominate over the main action and lead to erroneous view selection. Additionally, large flaws in silhouettes, for instance due to shadows, can harm view selection as well.

Real world data. To show the applicability of the proposed view selection method to more challenging videos we filmed our local basketball team during training using three fixed cameras. The cameras were set along the court, as shown on Figure 13 (a). In this scene, three players performed a drill that included running and free throws.

We extracted the players' silhouettes for each camera using simple background subtraction. This led to very noisy silhouettes. Furthermore, in significant parts of the scene the players were either very close to each other or occluded each other. Thus it was not practical to treat each player independently. Instead, we applied our view quality estimation to the joint shape of all three players, as if they were a single subject.

As illustrated in Figure 13 (b), camera 1 suffers from severe occlusions, camera 2 suffers from partial occlusions, while camera 3 captures most of the drill without occlusions. Our view quality rating reflects this nicely. These results demonstrate that the proposed view quality rating can be applied to single and multi-player scenes as one.

Figure 14 shows results of a single-player throwing a ball. Here our approach nicely detects the side view, where the throwing of the ball is best captured.

3D graphics data. Additionally to the videos taken by real video cameras, our viewpoint quality estimation is also relevant in 3D graphics, or specifically in 3D games. To show the applicability of the proposed method in this field we choose a scene of two hugging people from Sims 3 game, filmed from eight different angles. Note,

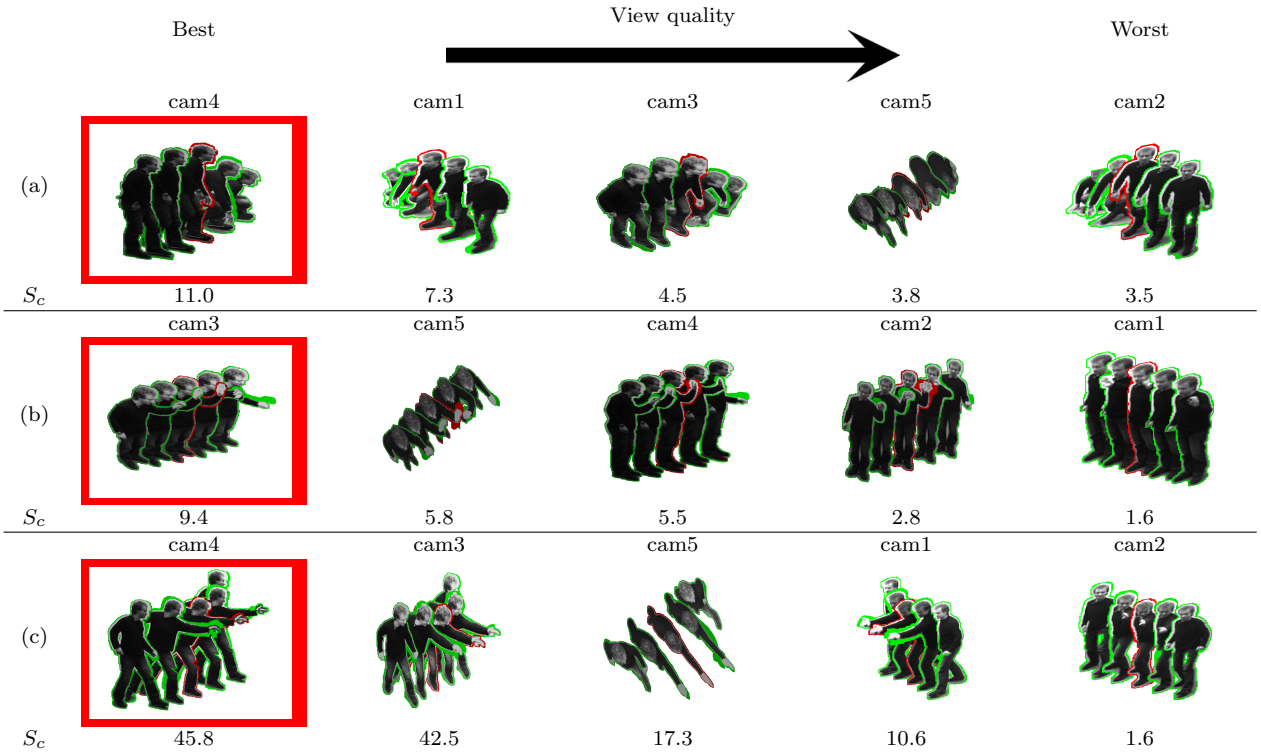


Fig. 11 View selection on IXMAS. Examples of view selection applied to IXMAS sequences of a single person. Top row (a) shows a "get-up" action, which is visible clearly from any angle, thus the view qualities do not vary too much. Rows (b) and (c) show actions where some views are preferred. In this case the different views are nicely rated according to the visibility they provide. Views showing clearly all the limbs are ranked best while views with severe self-occlusion are ranked worst.

that in this case perfect silhouettes are available. Since the figures touch each other most of the action, we treated them as a single subject and applied the viewpoint quality of Eq. (8). As shown in Figure 15, the side views get a higher ranking while the front and back views, with severe self occlusions, are least preferred. This matches what one typically expects to see in "hugging people" videos.

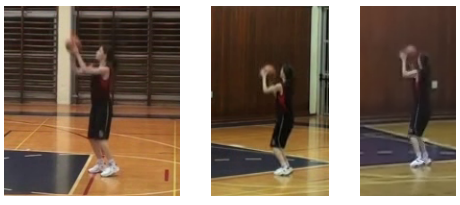


Fig. 14 A single player throwing a ball viewed from 3 viewpoints. As expected, the side view gets the highest rate.

6.2 Action recognition

As far as we know, there are no databases with ground-truth view selection, hence quantitative evaluation is somewhat difficult. To demonstrate that our technique selects good views we test its performance as a preprocessing step for action recognition. Given an action filmed from multiple angles we select a single view using two methods: (i) by selecting the view with maximum quality measure S_c of Equation (8), (ii) by selecting the view with the highest $\hat{s}$ of Equation (10). We then classify this view only. Our experiments show that selecting a single view can improve the recognition rates, implying that the views we select are good for recognition.

We test this framework on the IXMAS multi-view dataset (`IXMAS`). We split the long videos according to the provided ground-truth, into shorter action clips, so that each shows a single action. Following previous work we test on a leave-one-actor-out scenario, i.e., we take all the performances of one actor as the testing set and all other actors as the training set. We train a single classifier for all the available views, since the actors are oriented freely in the space. Having that classifier, we select the best and classify only it. In our experiments we used only 10 actors (excluding Srikumar and Pao) and 11 actions (excluding "throw" and "point") for fair comparison with previous work, which excluded these as well.

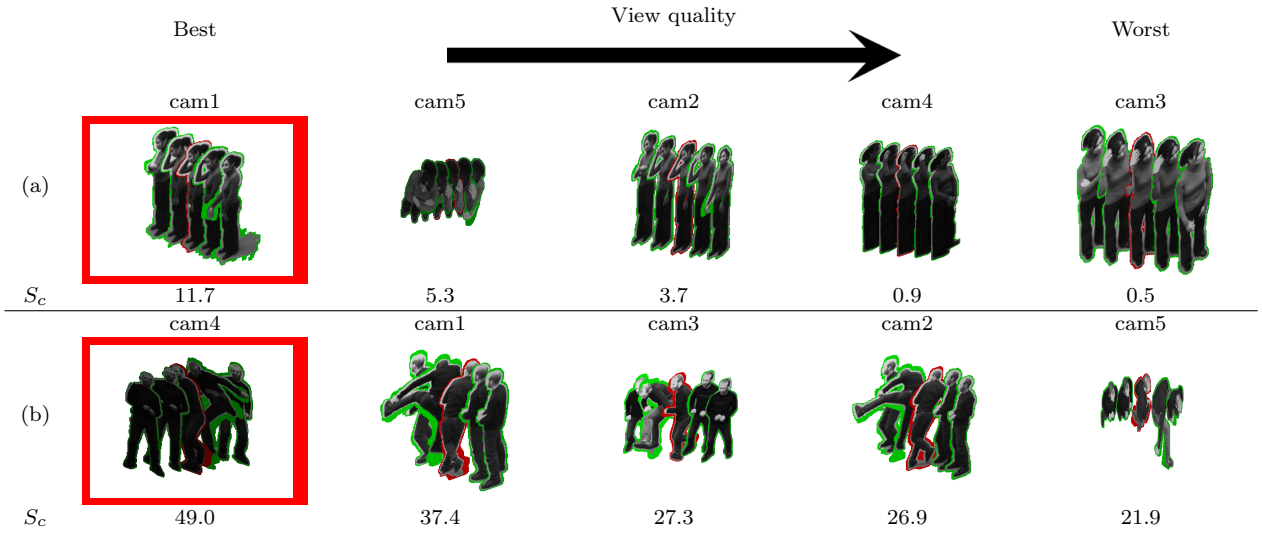


Fig. 12 Failure cases of view selection. (a) The action "scratch head" is barely visible from the selected view, however it was selected due to large changes in the attached shadows. (b) Camera 4 was selected for the "kick" action since there was a significant arm motion, which confused our measures.

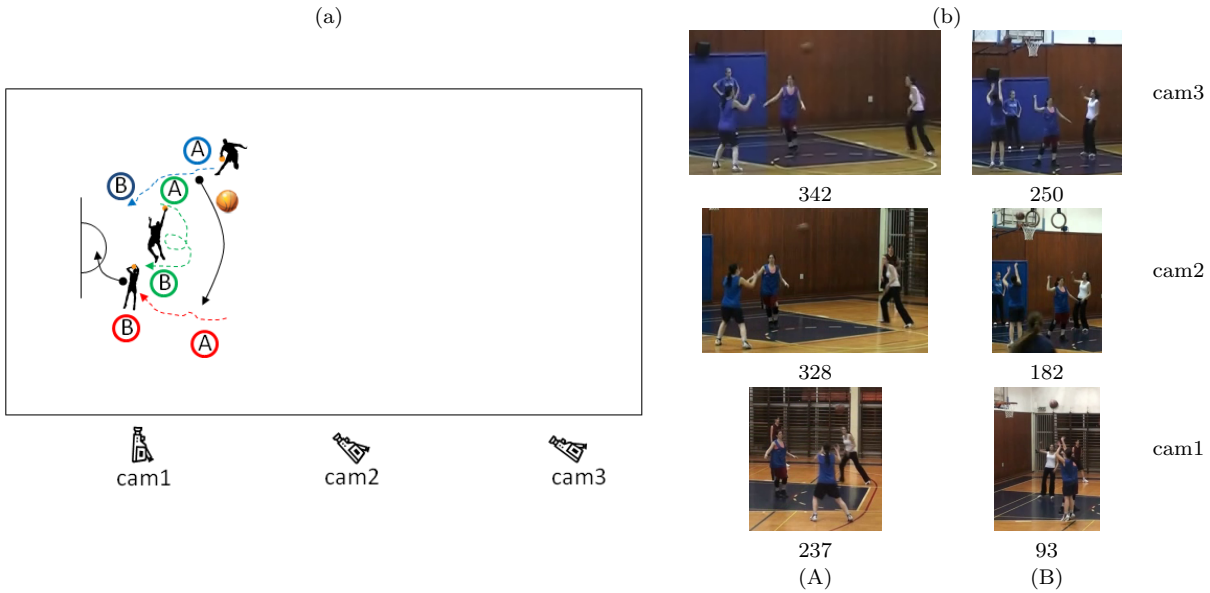


Fig. 13 (a) A schematic sketch of the filming setup of a basketball drill. The blue, green and red curves illustrate the paths of the players motion. The labels A and B mark the location in the field of each player at moments A and B. (b) Example frames from the three cameras at moments A and B depicted in the sketch. Camera 1 received the lowest rates since the arm motion in throwing the ball is occluded in (A) and the players occlude each other in (B). Camera 2 received higher rates since there are less self occlusions. Camera 3 got the highest view quality rates since there are no occlusions and the arm motion is clearly visible both in (A) and in (B). Please view the supplemental video.

In the first case, for each clip in the test set we evaluate the viewpoint quality provided by each camera using the proposed measure of Eq. (8). We then classify the action in the view with the highest score using one of three monocular recognition methods: (i) the silhouette based approach of Gorelick et al. (2007) [2], (ii) the view invariant approach of Junejo et al. (2010)[3], and (iii) the silhouette and optical flow based approach of Tran and Sorokin (2008)[4]. We further evaluate the quality of each clip using the measure of Equation (10),

[2] For (Gorelick et al., 2007) we obtained code from the authors.

[3] For (Junejo et al., 2010) we used our own implementation which obtains results similar to those reported in the original paper.

[4] For (Tran and Sorokin, 2008) we used authors' code available at their website with 1NN classifier. However, we used a slightly

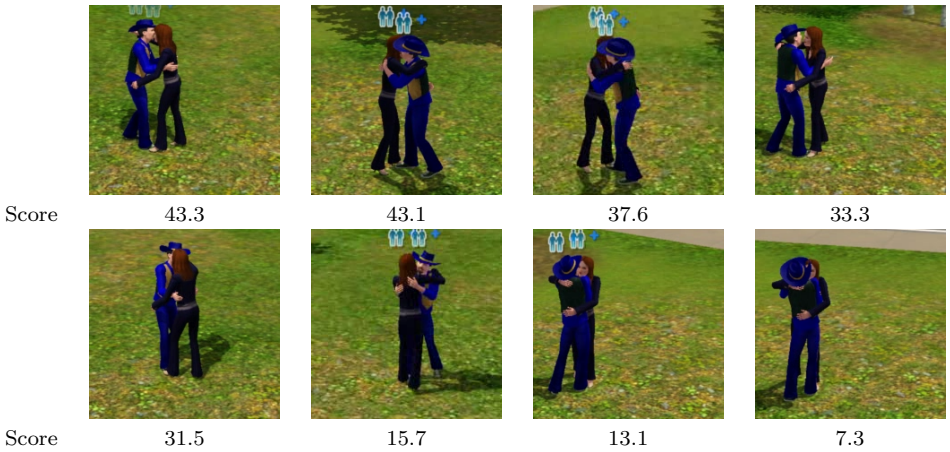


Fig. 15 View quality estimation for hug action in Sims 3 game. The different views are nicely rated according to the visibility they provide. Views where the interaction is clearly visible are ranked better than views in which one of the figures is occluded.

after training according to the same three recognition methods. For each method we select the best view and classify it.

We compare the results of the recognition after view selection with three other options: (i) average recognition rate, which reflects random selection of views, (ii) the rate of the "best" camera and (iii) the rate of the "worst" single camera. In "best" / "worst" camera we refer to the single camera with the highest / lowest recognition rate. In practice, selecting the "best" camera is not feasible, since it requires an oracle that a-priori tells us which of the fixed views will be better. However, this is the best rate that can be achieved from a single fixed camera. On contrary, a wrong selection of the camera could lead to "worst" camera rates.

Table 1 shows that the proposed view selection either matches, or improves the results of the "best" camera. This implies that the selected views are those where the action is recognizable, which satisfies the goal of this work. It is further interesting to note that the "best" fixed camera is different for each recognition method. This implies that on average different methods have preference for different viewpoints. Nevertheless, our view selection succeeds in detecting those views which are recognizable by all methods.

One of the limitations of the IXMAS dataset is that the cameras cover only half a circle around the action. Furthermore, although not instructed to do so, the actors naturally positioned to face one of the cameras, hence, there are no back views. To extend the data to include the back views as well we filmed a similar multiview action data-set[5] using the setup described in our golf-swing and dance experiments. For this dataset, the recognition rates using Gorelick's method were similar with and without view selection and stood on $\sim 50\%$. View selection had no benefit in terms of recognition rates, since the performance of the recognizer is poor. It is still beneficiary in the sense that recognition needs to be applied to a single view rather than multiple ones. We make this data available to the public at `http://cgm.technion.ac.il`.

different experimental setup, thus yielding slightly different results.

[5] The dataset is available at `http://cgm.technion.ac.il/Computer-Graphics-Multimedia/Resources/Resources.php`

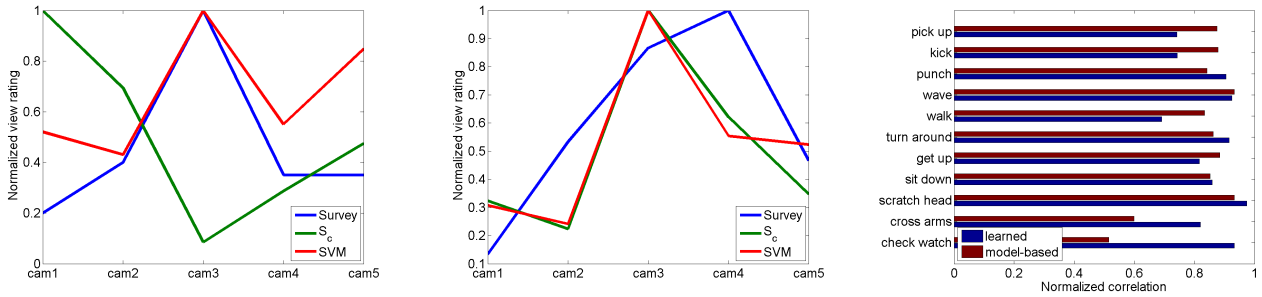
6.3 Comparison to human preference

Another experiment that could be useful for evaluation of the proposed view selection methods is comparison to user preferences. We performed a user study that involved 45 volonteers. Each participant was shown two random views of the same action and was asked to choose the prefered one. The chosen view scores 1 and the other scores 0. Each volonteer repeated the procedure for every action in the IXMAS dataset. After collecting all the data we sum up the scores for each action and for every view.

The comparison of the user preference to our automatic view selection is shown in Figure 16. For some actions, e.g, "check watch" (Figure 16 (a)) the correlation between the human preference and our results is limited, while for other actions it is high, e.g., "wave" (Figure 16 (b)). The correlation of each of our methods to the human preference is shown on Figure 16 (c). As can be seen, there is relatively high correlation in most of the actions. Our success is limited mostly for actions where silhouettes do not suffice.

Table 1 Comparison of recognition rates for different recognition methods shows that the proposed view selection performed before recognition often improves the best fixed camera rate. Note that an a-priori selection of the "best" camera is not possible. We mark in parentheses the label of the fixed camera that turned out to provide the best/worst recognition rates. Note that for each recognition method the best performance was obtained with a different camera.

	$\hat{s}$ Eq. (10)	S_c Eq. (8)	Fixed Camera		
			Average	Best (selected a-posteriori)	Worst (selected a-posteriori)
(Gorelick et al., 2007)	81	81	73	77 (cam4)	63 (cam5)
(Tran and Sorokin, 2008)	88	89	85	88 (cam3)	82 (cam5)
(Junejo et al., 2010)	65	65	64	67 (cam1)	57 (cam5)



"check watch" (a) "wave" (b) (c)

Fig. 16 Comparison of the proposed automatic view selection to human preference. (a) In the "check watch" action the users' preference is highly correlated to the learning method of view selection. On the other hand the model-based approach prefers other views. (b) In the "wave" action both methods prefer the same views as the human. (c) The normalized correlation between the human preferences and each of the methods, for all the actions of IXMAS. The correlation is rather high, suggesting our approach provides a good match to humans.

7 Robustness to the silhouette quality

The proposed view selection method fully relies on silhouettes of the human actor. Since the silhouette extraction process is never perfect we preprocess the silhouttes before the creation of the space time shape. First we dilate each silhouette using a disk structure element with size depending on the figure size. Then we leave only the biggest connected object which corresponds to the actor's body. An example of this process is shown in Figure 17. It can be seen that the preprocessing succesfully removes most of the unshmoothness of the space-time shape while preserving the parts that belong to large motions.

In our experiments we worked with different types of data. Some videos were filmed in relatively clean laboratory conditions (Figure 17 (a)) while others were filmed "in the wild" (Figure 17 (b)). After the preprocessing described above we managed to obtain space-time shapes of acceptable quality, even for the noisy data like the basketball training. Hence, our view selection methods performed well.

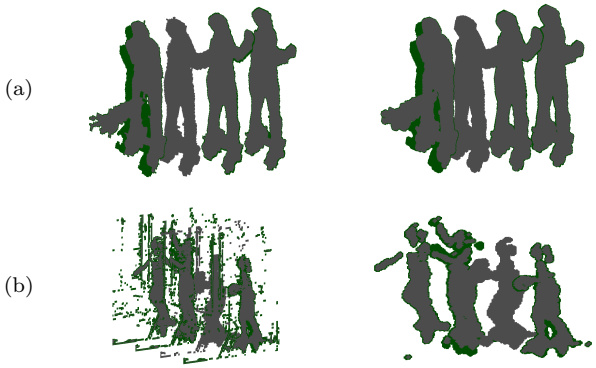


Fig. 17 The space-time shape for golf swing (a) and basketball throw (b) before and after the cleaning process. As one can see the shape before cleaning exibits noisy regions. This noise is removed during the cleaning and the resulting shape is much smoother, and suffices for view selection.

8 Conclusion

This paper presented a method for selection of the best viewpoint for human actions. To determine better views we compute a visibility score based on properties of the space-time shape induced by the actor's silhouettes. Additionally, we learn the better views according to

the performance of any given action recognizer. Our experiments show that the proposed approach can successfully estimate the action visibility provided by each camera. Such estimation can be used for automatic selection of a single best view of one or more actors. Furthermore, selecting the best view of the action prior to the recognition improves the rates of the monocular action recognition method, together with speeding them up (since we need to recognize only one view).

References

J. Assa, D. Cohen-Or, I.C. Yeh, and T.Y. Lee. Motion overview of human actions. In *International Conference on Computer Graphics and Interactive Techniques*. ACM New York, NY, USA, 2008.

J. Assa, L. Wolf, and D. Cohen-Or. The Virtual Director: a Correlation-Based Online Viewing of Human Motion. In *Eurographics*, 2010.

F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183–193, 1954.

L. Ballan, G.J. Brostow, J. Puwein, and M. Pollefeys. Unstructured video-based rendering: interactive exploration of casually captured videos. In *ACM SIGGRAPH 2010 papers*, pages 1–11. ACM, 2010.

Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

U. Bordoloi and H.W. Shen. View selection for volume rendering. In *IEEE Visualization*, volume 5, pages 487–494. Citeseer, 2005.

M. Christie, P. Olivier, and J.M. Normand. Camera control in computer graphics. In *Computer Graphics Forum*, volume 27, pages 2197–2218. Citeseer, 2008.

H. El-Alfy, D. Jacobs, and L. Davis. Assigning cameras to subjects in video surveillance systems. In *Proceedings of the 2009 IEEE international conference on Robotics and Automation*, pages 3623–3629. Institute of Electrical and Electronics Engineers Inc., The, 2009.

J. Feldman and M. Singh. Information along contours and object boundaries. *Psychological Review*, 112(1): 243–252, 2005.

L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

R. Goshorn, J. Goshorn, D. Goshorn, and H. Aghajan. Architecture for cluster-based automated surveillance network for detecting and tracking multiple persons. In *1st Int. Conf. on Distributed Smart Cameras (ICDSC)*, 2007.

`IXMAS`. http://charibdis.inrialpes.fr.

G. Johansson. Visual perception of biological motion and a model for its analysis. *Perceiving events and objects*, 1973.

I.N. Junejo, E. Dexter, I. Laptev, and P. Pérez. View-Independent Action Recognition from Temporal Self-Similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

G. Kindlmann, R. Whitaker, T. Tasdizen, and T. Moller. Curvature-based transfer functions for direct volume rendering: Methods and applications. In *Proceedings of the 14th IEEE Visualization 2003*, page 67. IEEE Computer Society, 2003.

N. Krahnstoever, T. Yu, S.N. Lim, K. Patwardhan, and P. Tu. Collaborative Real-Time Control of Active Cameras in Large Scale Surveillance Systems. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2*, number 2008, 2008.

C.H. Lee, A. Varshney, and D.W. Jacobs. Mesh saliency. *ACM Transactions on Graphics*, 24(3):659–666, 2005.

M. Mudge, N. Ryan, and R. Scopigno. Viewpoint quality and scene understanding. *Vast 2005*, page 67, 2005.

B. Schölkopf and A.J. Smola. *Learning with kernels*. the MIT Press, 2002.

D. Tran and A. Sorokin. Human activity recognition with metric learning. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, page 561. Springer-Verlag, 2008.

P.P. Vazquez, M. Feixas, M. Sbert, and W. Heidrich. Automatic view selection using viewpoint entropy and its application to image-based modelling. In *Computer Graphics Forum*, volume 22, pages 689–700. Citeseer, 2003.

T. Vieira, A. Bordignon, A. Peixoto, G. Tavares, H. Lopes, L. Velho, and T. Lewiner. Learning good views through intelligent galleries. In *Computer Graphics Forum*, volume 28, pages 717–726. John Wiley & Sons, 2009.