https://doi.org/10.2352/ISSN.2470-1173.2018.2.VIPC-205

Toward Automatic and Objective Evaluation of Synchronization in Synchronized Diving Video

Yixin Du and Xin Li; West Virginia University; Morgantown, WV 26506-6070, U.S.A.

Abstract

Most sports competitions are still judged by humans; the process of judging itself is not only skill and experience demanding but also at the risk of errors and unfairness. Advances in sensing and computing technologies have found successful applications to assist human judges with the refereeing process (e.g., the well-known Hawk-Eye system). Along this line of research, we propose to develop a computer vision (CV)-based objective synchronization scoring system for synchronized diving - a relatively young Olympic sport. In synchronized diving, subjective judgement is often difficult due to the rapidness of human motion, the limited viewing angles as well as the shortness of human memory, which inspires our development of an automatic and objective scoring system. Our CV-based scoring system consists of three components: (1) Background estimation using color and optical flow clues that can effectively segment the silhouette of both divers from the input video; (2) Feature extraction using histogram of oriented-gradients (HOG) and stick figures to obtain an abstract representation of each diver's posture invariant to body attributes (e.g., height and weight); (3) Synchronization evaluation by training a feed-forward neural network using cross-validation. We have tested the designed system on 22 diving video collected at 2012 London Olympic Games. Our experimental results have shown that CV-based approach can accurately produce synchronization scores that are close to the ones given by human judges with a MSE of as low as 0.24.

Introduction

The proliferation of sports game broadcasting calls for better ways of acquiring, processing, delivering and analyzing sports video contents. Fast advances in sensing and computing technologies have enabled the use of computer vision and video processing techniques to facilitate sports training, make sports video more entertaining and interactive. Some popular applications of sports video processing include video summarization [20], tactics and performance analysis [21], augmented reality presentation of sports [22], and referee assistance [5]. These applications have inspired many novel development of computer vision tools and video processing systems. For example, multi-camera replay system is often used to review the real-time action in basketball games; yellow-line technology has been widely adopted in the broadcasting of football games; augmented reality was used in the broadcasting of swimming video (e.g., highlighting the world and Olympic records) and alpine skiing video (e.g., overlaying of two athletes in space) to enhance the viewer's experience; and finally the well-known Hawk-Eye technology tracing a tennis ball's trajectory in order to ensure the fairness of tennis games.

Among hundreds of events in summer Olympics, synchronized diving is relatively young being adopted as an Olympic sport in 2000. Synchronized diving requires both perfect synchronization and execution performance from two divers. According to the diving officials manual released by Federation Internationale de Natation (FINA), there are usually seven to nine judges among whom three or five of them mark the synchronization [6] and the rest the execution. However, judges sit on the two sides of divers, and therefore their judgement is only based on side-view of divers which could possibly lead to bias. Meantime, the judges remain still while two divers' altitude keeps changing, which makes it difficult if not impossible to align the judges' line-of-sight with both divers. Since the event usually lasts for a few seconds only, judging the synchronization of two divers out of memory also has the tendency of making errors. In summary, both fixed viewpoint and rapid motion of divers potentially contribute to potential bias or errors in human's judging process. These observations motivate us to design a computer vision based system which takes a front-view video as the input and outputs a synchronization score to facilitate the referring process (so human judges need to focus on execution performance only). Such an automatic and objective synchronization scoring system is expected to improve the fairness and accuracy of refereeing process for synchronized diving.

There are primarily three technical challenges during the designing of a computer vision based synchronization scoring system. First, accurate segmentation of both divers' silhouette from the input video is error-prone due to complex background and rapid camera motion. Segmentation becomes even more difficult when divers enter the water (note that this timing matters to the scoring on synchronization). Second, it is non-trivial to choose appropriate feature representation for human motion attributes invariant to human physiology such as body height and weight. In other words, judging the synchronization of motion between two athletes is often more challenging as their body attribute difference increases (for this reason, athletes of similar height and weight are often preferred). Finally, in the situation of lacking synchronization (e.g., varying altitude v.s. varying speed), how to *objectively* measure the severity of synchronization lacking is a nontrivial issue. We note that sometimes specific point deduction rules are always articulated even in the manual of diving officials released by FINA.

To tackle those technical challenges, we propose to develop an automatic and objective scoring system consisting of the following three components (as shown in Figure 1): (1) Background estimation using color and optical flow clues that can effectively segment the silhouette of both divers from the input video. The novelty of our approach lies in the fusion of motion and color related priors (e.g., rapid object motion and human skin color) for diving video; (2) Feature extraction using histogram of oriented-gradient (HOG) [23] and stick figures [24] to obtain an abstract representation of each diver's posture. The former has been

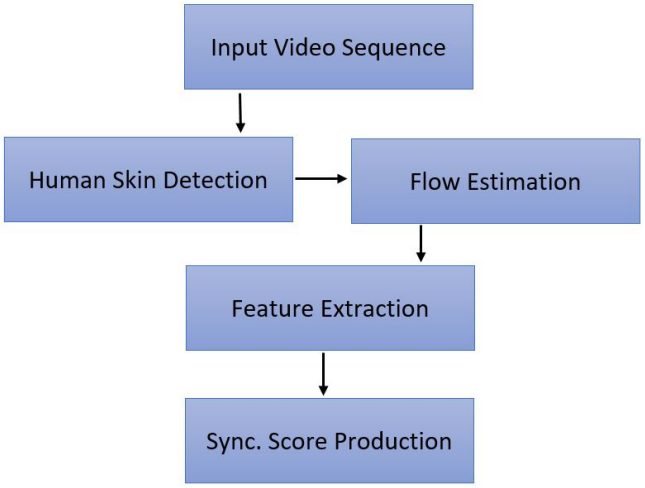


Figure 1: The components the proposed approach. Divers' silhouette is extracted from an input video using a color and optical flow combined approach. Then, we extract features from the segmentation results, and use the features to train a neural network to produce synchronization scores.

widely used for detecting humans in images and video; while the latter is adopted for their invariance to body attributes (e.g., height and weight); (3) Synchronization evaluation by training a feed-forward neural network using cross-validation. The method of cross-validation (i.e., "test" the model in the training phase) is used for the purpose of avoiding over-fitting (a common problem in the practice of neural networks). We have tested the designed system on 22 diving video collected at 2012 London Olympic Games. Our experimental results have shown that CV-based approach can accurately produce synchronization scores that are close to the ones given by human judges with a *MSE* of as low as 0.24.

Literature

In this section, we briefly review the literature of sports video processing and its applications including video abstracting/summarization, tactics analysis, and computer-aided referee assistance.

Video abstracting/summarization is a technique of distilling the essence of a long video into a concise and representative key frames [10], [4]. When applied to sports videos, it provides audience with the highlights of a game so they do not need to watch the complete video for the sake of time-saving [11]. Game highlights often include scoring events in basketball and soccer, touchdown in football games, breaking world record moments in swimming, etc. Since it is difficult to detect game highlights with video/image data only , many works in this area have adopted the strategy of working with audio features to make it more tractable (e.g., [12][13][14]).

Tactics analysis refers to automatic understanding of strategic tactics adopted by athlete(s) from video data. Coaches and players are often interested in tactics analysis because it offers a supplementary tool for assisting athlete training and improving individual or team players' performance. For individual training such as tennis and swimming, video-based tactics analysis can analyze player's movement pattern (e.g., serving in tennis and different strokes in swimming), extract useful information (such as the trajectory of a ball [15][16]) and facilitate the communication between athletes and coaches. In team sports [17], [18] video-based tactics analysis focus more on the collective movement patterns (e.g., the position and trajectory of basketball/soccer players) of both teams, which can be exploited by the coaches to adjust the offensive or defensive strategies during the practice.

In such sports such as diving and gymnastics, human judges are involved to score the performance of athletes. Even though it takes experience and skills for the qualification of serving as judges, the actual fairness of scoring by human judges is often questionable due to human errors and bias. Computer-based refereeing could facilitate the works of human referee or judges [19] - in some sports such as basketball, video replay has been adopted to facilitate referees to have a second opinion when real-time decision raises doubt due to various uncertainty factors (e.g., poor view angle or blocked view due to occlusion); in other sports such as tennis, Hawk-eye technology has become mature enough to become the ultimate judge in the case of disagreement. It is reasonable to expect that computer-based refereeing could play an even more important role in the future as sensing and computing technologies evolve.

Video-based synchronization analysis for diving has been studied before but scarcely (the only reference we can find in the open literature is [7]). In that work, the divers' silhouette is first extracted using foreground segmentation, background reconstruction, and silhouette detection. A five-dimensional feature vector is constructed for each input video based on the FINA diving rules [6], including the take-off height similarity, the coordinated timing of motion, the similarity of angles of entry, the comparative distance from the springboard of entry, and the coordinated timing of entries. Rankboost algorithm [3] is used to evaluate the model's performance. One major limitation of [7] is that the construction of feature vectors requires heavy manual intervention. For example, the similarity of angels of entry is approximated by water sprays which is achieved by manually labeling a bounding box around the spray, which is often a time-consuming process. Based on those observations, we propose to develop a fully automatic and objective scoring system without manual intervention.

Methodology

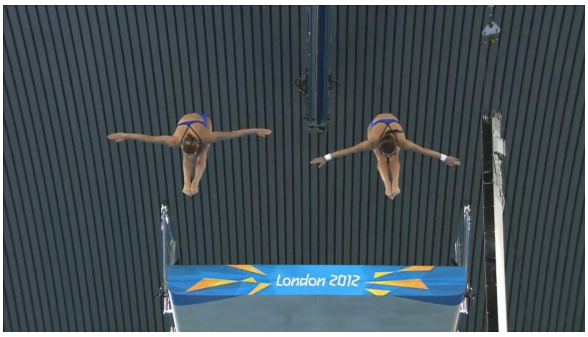
Our system consists of three components: video segmentation, feature extraction and synchronization score production. We will elaborate the design of each component in this section focusing on the incorporation of a priori knowledge related to the event (synchronized diving).

Silhouette Extraction using Optical Flow and Color

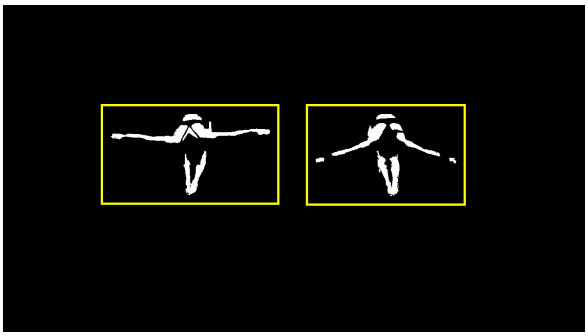
The challenge with robust segmentation of diving video lies in sophisticated object motion and complex background. Here we propose to develop a context-aware diving video segmentation technique by jointly exploit *motion* and *color* information. On the motion part, we have adopted optical flow estimation [2] - a widely used technique to extract motion information from video sequences. Based on estimated optical flow, video sequence can be segmented into foreground (moving objects) and background

(still scenes) layers. However, inaccurate segmentation often occurs due to either incomplete object boundaries or errors in optical flow estimation [8]. In the scenario of synchronized diving, all videos contain rapid camera motion, which is a primary reason for causing errors in optical flow estimation (the faster an object moves, the more difficult it becomes to establish the correspondence between adjacent frames). Moreover, in the case of diving, deformable motion arising from rapid variation of body shape makes it more challenging to reliably separate the foreground (divers' silhouette) from a moving background (due to camera motion).

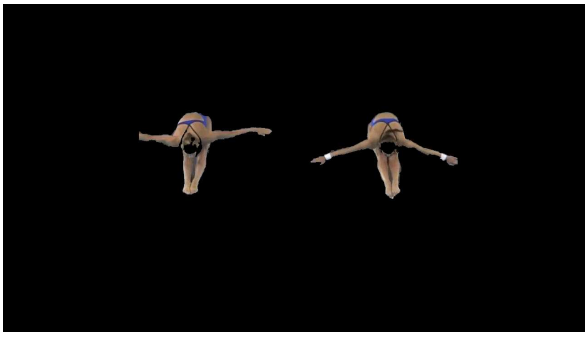
In view of the difficulties with using motion clue alone, we propose to exploit color information to facilitate the segmentation process. More specifically, one strong prior with all diving videos is that a significant portion of the diver's skin is visible and skin color is relatively consistent. Detecting human using skin color [1] is a widely studied topic and here this technique is adopted to obtain a good initial estimate of foreground (divers). Among various human skin detectors using color, we have empirically tested both HSV and Lab color spaces. Our experiments have shown that under the context of diving video segmentation, Lab space is slightly superior when compared with HSV space. So we opt to perform skin detection via hard thresholding in Lab space and obtain an initial foreground estimation. Then the background is estimated by optical flow method based on Papazoglou and Ferrari [8]. Finally, the outcomes of skin detection and optical flow estimation are combined together, producing the silhouette extraction result. Figure 2 illustrates the segmentation technique in our work, in which (a) is the original frame of an input video, (b) is human skin detection in Lab color space, and (c) is the final silhouette extracted by combining optical flow and skin detection.



(a)



(b)



(c)

Figure 2: (a) Original frame of an input video. (b) Human skin detection in Lab color space. (c) Final silhouette extracted by combining optical flow and skin detection.

Feature Extraction

With extracted silhouette, the natural next step is to find a suitable feature representation for analyzing the synchronization between two divers. Note that the task of judging the synchronization of motion between two athletes is often more challenging as their body attribute difference increases even for human judges. For this reason, athletes of similar height and weight are often preferred even during the selection and training by diving coaches. Ideally we would like to pursue feature representations that are invariant to body weight and height; but at the same time we also want to preserve motion-related information since it is critical to the objective assessment of synchronization. In [7], five set of features are extracted from diving video based on the diving rules published by FINA; we argue that such model-based approach has its limitations because the interpretation of rules often involves ambiguity (e.g., the rules only specify what are important factors to consider but do not articulate objective procedures for calculating the penalty or synchronization score).

In this paper, we advocate a learning-based approach where only discriminative features are extracted to support the production of synchronization score by a neural network. In other words, we do not explicitly extract synchronization-related features like [7] but target at feature representations suitable as inputs to the network of synchronization analysis (the last component). Two sets of features are considered here: histogram of oriented gradients (HOG) [9] and stick figures [24] (please refer to Fig. 3). The former - with proper normalization by the patch size - is approximately invariant to the height of a human; and the latter - that has been widely used in human motion analysis such as gymnastics and gaming - is invariant to the body weight. Other feature representations (e.g., contour-based and volumetric [25]) are deemed less appropriate for the analysis of motion synchronization.

Histogram of oriented gradients (HOG) is a feature descriptor that has been particularly effective for detecting humans in an image. The concept of HOG was firstly described in 1984 but it only became widespread since the publication of [9] in 2005. First, the gradient computation is achieved by applying two filter kernels ($[-1,0,1]$ and $[-1,0,1]^T$) (discrete implementation of directional derivatives) to an image. Then, the cell histogram is created by counting the frequency within each orientation bin, followed by grouping the cells into larger spatially connected blocks

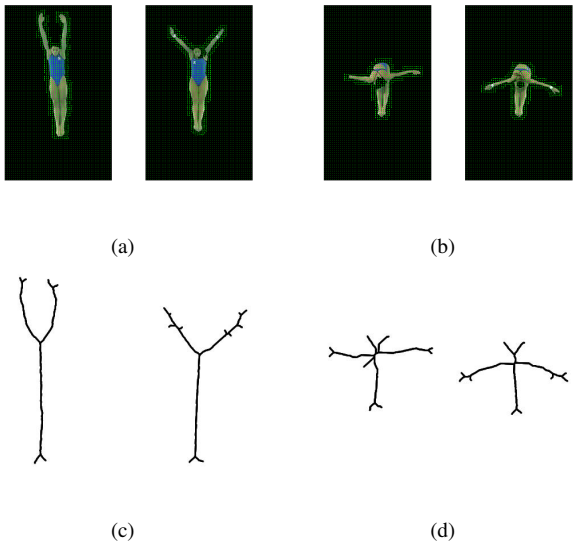


Figure 3: Top row: HOG feature extraction. The two divers' movement in (a) differs more compared to the ones in (b). Numerically, the HOG matching error score is 0.6 in (a) and 0.25 in (b), the lower the score, the better the synchronization. Bottom row shows the stick figures extracted, where the Hausdorff Distance is 0.77 in (c) vs 0.23 in (d), the lower the distance, the better the synchronization.

and block normalization. We choose HOG because it is not only really simple to compute but also highly descriptive - i.e., abstraction of the full body profile of each diver as opposed to other local feature descriptors such as SIFT. Another advantage is that HOG can produce a fixed-length feature vector under normalization (regardless of varying heights), allowing us to easily compute the Euclidean distance between HOG-based feature vectors. Additionally, note that the difference in body weights could contribute to the Euclidean distance of the HOG feature vectors. To take this into account, we propose to use stick figures for assisting the characterization of each diver's movement which is invariant to body weights. The skeleton image of each diver is extracted using some morphological operations, followed by calculating the Hausdorff Distance (HD) between two stick figures.

$$\mathrm{HD}(P,Q) = max_{p\in P}\{min_{q\in Q}\{\|p-q\|\}\}. \quad (1)$$

Synchronization Score Production

With two feature vectors $[d_1, d_2]^T$ for each frame of input video, where d_1 is the Euclidean distance of the HOG feature and d_2 is the Hausdorff Distance of the stick figures, we can cast the task of synchronization score production as a regression problem. Solving a regression problem usually takes a parametric or non-parametric approach. Building a parametric regression model requires the knowledge of the observation process (i.e., how data are acquired?) and usually works better for data in low-dimensional space (e.g., the five-dimensional feature space used in [7]). With some training data available, it is possible to obtain a strong ranking function by combining multiple weak ones - e.g., the RankBoost [3] approach that has been adopted by [7].

By contrast, a non-parametric regression model such as the neural network do not require explicit modeling of the observation system (but still assuming synchronization scores marked by human judges are available as training data). Instead, what a neural network attempts to learn is a nonlinear mapping from input space (feature vectors) to output space (synchronization scores). It is particularly suitable for machine learning tasks with hand-crafted features such as HOG and stick figures here. We do recognize the possibility of training a deep convolutional neural network for automatically learning the feature and producing the synchronization score but it is beyond the scope of this work.

In this paper, we have adopted a simple feedforward neural network without any cycle among the connections of neurons. The information simply moves in one direction (feed-forward): from input layer, to hidden layer, and finally output layer. We choose a two-layer neural network with sigmoid hidden neurons and linear output neurons. The neural network is trained with the Levenberg-Marquardt backpropagation algorithm. The number of hidden neurons in hidden layer is the key parameter setting to prevent both underfitting and overfitting, and provide a good approximation to the input data. More implementation details can be found in the next section of experimental results.

Experiments

Dataset Preparation

We have collected a benchmark dataset containing 22 videos of men's and women's synchronized 10 meters platform diving from London 2012 Olympic Games. Each video lasts around 4-6 seconds (120-180 frames at 30fps) with a spatial resolution of 1280×720. All videos are front-view instead of side-view, which is more suitable for video-based feature extraction and synchronization scoring (note that our CV-based approach is supplementary to existing human-based since judges are all seated side-view). Table 1 is a summary of the dataset. The first two columns are video IDs and the number of frames; other columns are synchronization scores marked by human judges. The scores are from a $[0,10]$ interval. The average of five synchronization scores are used as the value of the response variable during training.

Silhouette Extraction

We report the segmentation results as described in the last section. In the first step, hard thresholding is performed in Lab color space in order to detect human skins. After some morphological operations to get rid of small false positive pixels, we draw a bounding box on each diver. This bounding box serves as an initial estimate of the FG. In the second step, the video is segmented using optical flow, and the major inaccuracy in the results are the false positive pixels that are labeled incorrectly in flow estimation. Finally, we combine the optical flow segmentation result with the skin detection result. Figure 4 illustrates the silhouette extraction without and with skin detection. It can be observed that skin detection does improve the accuracy of silhouette extraction.

Feature Extraction

The HOG and stick figure features are extracted from segmentation results for both divers. Intuitively, the more similar of the divers' movement, the lower the distances of HOG and stick figure features. Figure 3 shows some examples of feature extraction. The top row is HOG feature extraction: the two divers'

Table 1: Summary of the benchmark dataset: the first two columns are video IDs and the number of frames; the remaining columns are synchronization scores marked by human judges.

ID	Frames	S1	S2	S3	S4	S5
1	120	9	8.5	9	9	8.5
2	187	9	9	9	9.5	9
3	116	8.5	9	8.5	8.5	8.5
4	146	8.5	8	8.5	8.5	9
5	212	8	8	8.5	7.5	8
6	111	7.5	7.5	7	7.5	7.5
7	228	9.5	8.5	9.5	10	9.5
8	106	9	9	8.5	8.5	9
9	126	9	9	8	8.5	8.5
10	94	9.5	9.5	9.5	9	9
11	146	8.5	9	9	8.5	8.5
12	157	8.5	9	8.5	9	8.5
13	111	8.5	8.5	8	8	7.5
14	117	9	9.5	9.5	9	9.5
15	117	7.5	7.5	7	7.5	8
16	113	7.5	7.5	7	7	7.5
17	157	9	8.5	8.5	8.5	9
18	117	8.5	8.5	8	8	7.5
19	114	8	8	8	8	8.5
20	164	9	8.5	7.5	8.5	8
21	125	9	9	8.5	9	8
22	129	8.5	8	8.5	9	8.5

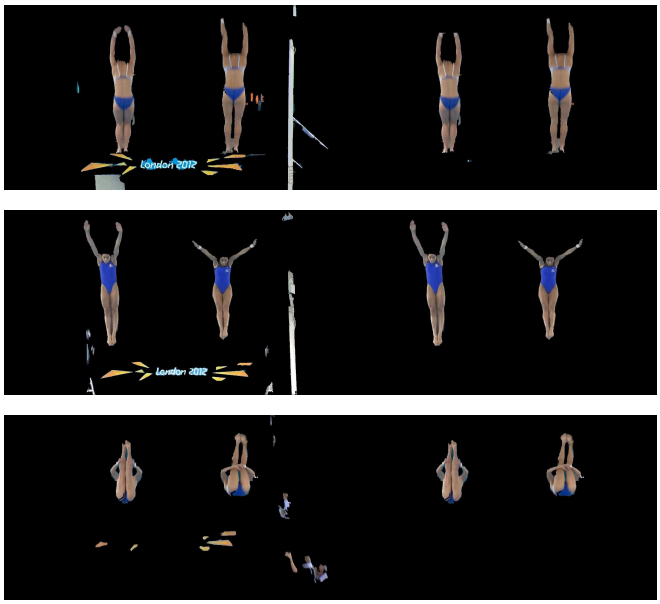


Figure 4: Silhouette extraction without (left) and with (right) human skin detection.

movement in (a) differs more compared to the ones in (b). Numerically, the HOG matching error score is 0.6 in (a) and 0.25 in (b), the lower the score, the better the synchronization. Bottom row shows the stick figures extracted, where the Hausdorff Distance is 0.77 in (c) vs 0.23 in (d), the lower the distance, the better the synchronization. For each frame, the HOG is represented using a fixed size vector for better estimating the Euclidean distances. Similarly, the Hausdorff Distance is calculated between two stick figures. There are 3013 frames in total, and we eliminated some frames which distances are too large. These large distances are mainly caused by segmentation inaccuracy. Finally, there are 2791 frames selected for predicting the synchronization scores.

Synchronization Score Production

In order to produce synchronization scores, we have trained a two-layer feed-forward neural network. In statistics and machine learning, overfitting happens when the model is too complex and there is not sufficient training data to describe the model; while underfitting means the model is too simple to capture the underlying characteristics of data. For feed-forward neural networks, the parameter which might cause overfitting or underfitting is the number of hidden neurons in the hidden layer. Thus, we have conducted experiments using 1 to 10 hidden neurons respectively, each of which has a 10-fold cross-validation with different training, validation and testing feature instances. Mean squared error (*MSE*) in the testing phase is used as performance measure. We have found the average *MSE* is the lowest (0.24) when using 6 hidden neurons.

The trained network is then used to produce the synchronization score on each of the 22 input video sequences. There is no overlapping in the training, validation, and testing data. After predicting the synchronization score for each frame, we compute the score of the video by averaging. Table 2 is the comparison of ground truth and the predicted synchronization score. The ground truth of one video is the average of the five synchronization scores listed in Table 1. The predicted value is generated using a neural network with 6 hidden neurons.

In order to illustrate the benefit of using both HOG and stick figure as a representation of diver's posture invariant to body weight, we conduct another two sets of experiments by training neural networks and predicting synchronization score using one class of feature only (either HOG or stick figure feature). Figure 5 is the comparison of prediction errors using networks trained with both HOG and stick figure features (in green), HOG feature only (in blue) and stick figure feature only (in red) respectively. It can be observed that 1) the best performance is achieved using both HOG and stick figure features; 2) HOG feature is better compared with stick figure features since overall the blue line's deviates less from zero than the red line, which shows that HOG feature is stronger and better than stick figure feature.

Failure Cases Analysis

Intuitively, the main source of error comes from the segmentation step. Here we opt to show some failure cases in segmentation, explain why those happen, and point out the direction of our future research that could lead to further improvements.

Figure 6 (a) illustrates the situation when one of the divers' body part is missing. We noticed that this type of situation contributes to the majority of mislabeled pixels in segmentation. The reason is that due to rapid camera motion and busy background, it's hard for flow estimation to correctly segment the two divers simultaneously with one single threshold setting. This problem happens quite often when one deals with multi-object or multi-

Table 2: The comparison of ground truth and the predicted synchronization score. The ground truth of one video is the average of the five synchronization scores listed in Table 1. The predicted value is generated using a neural network with 4 hidden neurons.

ID	Human-based	CV-based (This work)
1	8.8	8.911
2	9.1	8.983
3	8.6	8.694
4	8.5	8.676
5	8.0	7.88
6	7.4	7.519
7	9.4	9.269
8	8.8	8.693
9	8.6	8.721
10	9.3	9.414
11	8.7	8.915
12	8.7	8.84
13	8.1	8.285
14	9.3	9.363
15	7.5	7.422
16	7.3	7.19
17	8.7	8.625
18	8.1	8.179
19	8.1	8.294
20	8.3	8.186
21	8.7	8.847
22	8.5	8.286

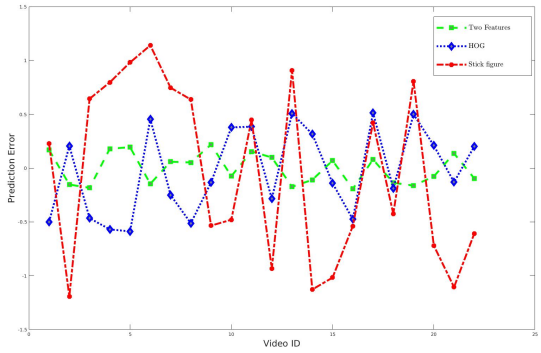
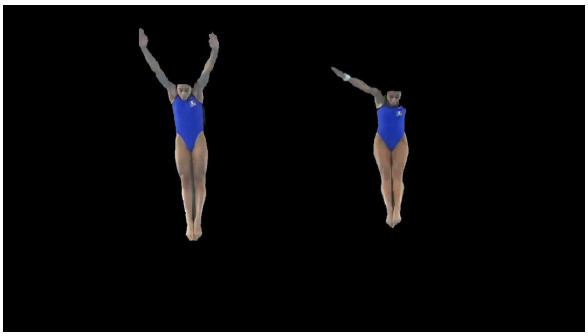


Figure 5: Comparison of prediction errors using networks trained with HOG and stick figure features (in green), HOG feature (in blue) and stick figure feature (in red) respectively. It shows that the best performance is achieved using both HOG and stick figure features, and HOG feature is more discriminative compared with stick figure features since overall the blue line's deviates less from zero than the red line.

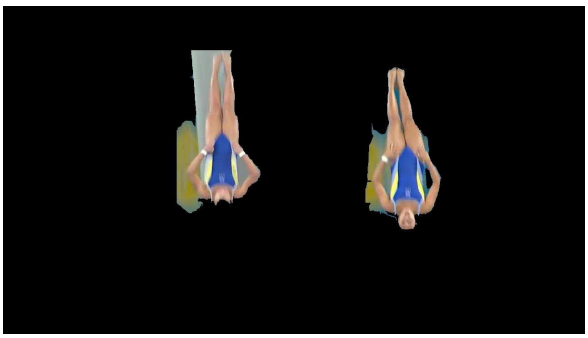
frame cases as there is no single optimal setting capable of dealing with all cases.

Figure 6 (b) shows the situation when the background pixels on the right hind side is falsely segmented. Since we use human skin detection from color, when the background pixels have similar color as skin color, or when some people other than the divers exist in the background, skin detection may fail. One possible



(a)

(b)



(c)

Figure 6: (a) Missing body parts due to sophisticated body and camera motions. (b) In skin detection using color, the background pixels are false classified as the their color is very close to human skin's color. (c) The false positive pixels in the background lie within the bounding box created in skin detection.

solution is to establish connections between flow estimation and color segmentation such that beside color cue, color segmentation also takes motion into account. The resulting reported pixels not only satisfy skin color constraint but also meet FG motion requirement.

Figure 6 (c) happens when false positive pixels in the background lie within the bounding box created in skin detection. This can be improved if the above two situations are effectively handled.

Conclusion

We propose a computer vision system that helps produce the synchronization score of synchronized diving. The input of our

system is based on the front-view diving videos. There are three steps in the proposed approach: silhouette extraction, feature extraction, and synchronization score prediction. Different from the past work on synchronization evaluation, our approach is fully automatic without the need of manual intervention. Moreover, it objectively evaluates the synchronization in a frame by frame basis, thus, avoiding subjective bias of human judges due to memory shortages. The system is tested on our benchmark dataset, and the experimental results have shown that the predicted synchronization score is very close to human judge's mark with very low prediction error.

References

[1] Jones, Michael J., and James M. Rehg. "Statistical color models with application to skin detection." International Journal of Computer Vision 46.1 (2002): 81-96.

[2] Liu, Ce, Jenny Yuen, and Antonio Torralba. "Sift flow: Dense correspondence across scenes and its applications." IEEE transactions on pattern analysis and machine intelligence 33.5 (2011): 978-994.

[3] Freund, Yoav, et al. "An efficient boosting algorithm for combining preferences." Journal of Machine Learning Research 4.Nov (2003): 933-969.

[4] DeMenthon, Daniel, Vikrant Kobla, and David Doermann. "Video summarization by curve simplification." Proceedings of the sixth ACM international conference on Multimedia. ACM, 1998.

[5] Yu, Xinguo, and Dirk Farin. "Current and emerging topics in sports video processing." Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005.

[6] http://www.fina.org/content/diving-rules.

[7] Ding, Haoyang, et al. "Synchronization analysis for synchronized diving videos." Multimedia and Expo, 2008 IEEE International Conference on. IEEE, 2008.

[8] Papazoglou, Anestis, and Vittorio Ferrari. "Fast object segmentation in unconstrained video." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[9] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[10] Lienhart, Rainer, Silvia Pfeiffer, and Wolfgang Effelsberg. "Video abstracting." Communications of the ACM 40.12 (1997): 54-62.

[11] Wang, Jenny R., and Nandan Parameswaran. "Survey of sports video analysis: research issues and applications." Proceedings of the Pan-Sydney area workshop on Visual information processing. Australian Computer Society, Inc., 2004.

[12] Xiong, Zixiang, Regunathan Radhakrishnan, and Ajay Divakaran. "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework." Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on. Vol. 1. IEEE, 2003.

[13] Xiong, Ziyou, et al. "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework." Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. Vol. 5. IEEE, 2003.

[14] Radhakrishan, Regunathan, et al. "Generation of sports highlights using a combination of supervised & unsupervised learning in audio domain." Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. Vol. 2. IEEE, 2003.

[15] Zhu, Guangyu, et al. "Trajectory based event tactics analysis in broadcast sports video." Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007.

[16] Yu, Xinguo, et al. "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video." Proceedings of the eleventh ACM international conference on Multimedia. ACM, 2003.

[17] Yu, Xinguo, et al. "Team possession analysis for broadcast soccer video based on ball trajectory." Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. Vol. 3. IEEE, 2003.

[18] Taki, Tsuyoshi, Jun-ichi Hasegawa, and Teruo Fukumura. "Development of motion analysis system for quantitative evaluation of teamwork in soccer games." Image Processing, 1996. Proceedings., International Conference on. Vol. 3. IEEE, 1996.

[19] Lam, M., et al. "Computer-assisted off-side detection in soccer matches." Proceedings of Technical Report, School of Information Technologies, University of Sydney (2003).

[20] Ma, Yu-Fei, et al. "A user attention model for video summarization." Proceedings of the tenth ACM international conference on Multimedia. ACM, 2002.

[21] Zhu, Guangyu, et al. "Trajectory based event tactics analysis in broadcast sports video." Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007.

[22] Han, Jungong, and Dirk Farin. "A real-time augmented-reality system for sports broadcast video enhancement." Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007.

[23] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2005.

[24] Lee, Hsi-Jian, and Zen Chen. "Determination of 3D human body postures from a single view." Computer Vision, Graphics, and Image Processing 30.2 (1985): 148-168.

[25] Aggarwal, Jake K., and Quin Cai. "Human motion analysis: A review." Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE. IEEE, 1997.

Author Biography

Yixin Du received the B.S. degree in Thermal Energy and Power Engineering from Tianjin University of Technology in 2012, and M.S. degree in Industrial Engineering from West Virginia University in 2014. He is currently a Ph.D. student in Department of Computer Science and Electrical Engineering at West Virginia University. He is working as a research assistant on algorithm design in computer vision and machine learning.

Xin Li received the B.S. degree with highest honors in electronic engineering and information science from University of Science and Technology of China, Hefei, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 2000. He was a Member of Technical Staff with Sharp Laboratories of America, Camas, WA from Aug. 2000 to Dec. 2002. Since Jan. 2003, he has been a faculty member in Lane Department of Computer Science and Electrical Engineering. His research interests include image/video processing, computer vision, biometrics and information security. Prof. Xin Li is an Fellow of IEEE.