17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Selecting an Iconic Pose From an Action Video

Geethu Miriam Jacob Björn Stenger
Rakuten Institute of Technology, Rakuten Group, Inc.
{geethu.jacob, bjorn.stenger}@rakuten.com

Abstract

This paper presents a method for selecting an iconic pose frame from an action video. An iconic pose frame is a frame showing a representative pose, distinct from other actions. We first extract a diverse set of keyframes from the video using unsupervised video summarization. A classification loss ensures that the selected frames retain high action classification accuracy. To find iconic poses, we introduce two loss terms, an Extreme Pose Loss, encouraging selecting poses far from the mean pose, and a Frame Contrastive Loss, which encourages poses from the same action to be similar. In a user preference study on UCF-101 videos we show that the automatically selected iconic pose keyframes are preferred to manually selected ones in 48% of cases.

1 Introduction

This paper addresses the task of selecting a single representative frame from a video, which lets people easily recognize the action class. This enables practical applications, such as automatically selecting a video thumbnail, or a good still shot from a sports video. We refer to this problem as *iconic pose keyframe selection*, and formalize it by selecting a frame which (i) summarizes the action well, and (ii) distinguishes it from other types of actions. Figure 1 shows examples for different types of actions. The figure shows a set of keyframes from UCF-101 [21] action videos *Kick*, *HighJump*, and *Bowling* and the selected iconic pose keyframe for each. We observe that these poses can be characterized by physical contact (Kick), extreme points of motion (HighJump) or by releasing an object (Bowling). This difference in semantics highlights the difficulty of the task. We observe that iconic poses tend to be extreme poses, far from an average or neutral pose.

Iconic pose keyframe selection can be seen as a single-frame version of keyframe detection for video summarization, where the selected frame is representative of the action. Prior work for keyframe selection used optical flow patterns [13, 23] or local features [7, 15]. Several methods that extract multiple frames for video summarization aim to maximize frame diversity [1, 2, 6, 12, 16, 17, 18]. Keyframe selection methods are typically unsupervised, however supervised methods include [7], where keyframes are manually selected, and [25], where frames for different action classes are selected using linear discriminant analysis (LDA). Our method is weakly supervised by providing action labels for whole video clips at the training stage and using an action classification loss term. Prior work in thumbnail selection typically ranks images based on an aesthetics score or the correlation of image to a text query [5, 11, 14, 20]. In this work we focus on videos showing activities and explicitly take body pose into account. Our method uses a novel loss function with two terms, in order to select poses that are both action-specific and extreme. We employ a keyframe extraction module [18] and an action classification module, trained on action labels as a weak supervision signal. The class labels of the actions give higher likelihood to frames relevant to a particular action class, while avoiding the selection of poses that could be confused with a different action class. The keyframe selection module in [18] uses two loss terms, reconstruction loss and diversity loss. In addition, we introduce two new loss terms, Frame Contrastive Loss (FCL) and Extreme Pose Loss (EPL), motivated by the aim to select similar poses for the same action, as well as selecting extreme poses, far from the mean pose.

Figure 1: **Iconic Pose Selection.** Each row shows five keyframes of a UCF-101 video. Frames with red border show iconic pose frames selected by the proposed method.

Our contributions include (1) a novel problem setting of finding a single iconic pose frame that represents the action in a video, (2) introducing a novel loss function for extracting the most representative and extreme human poses in the video using Procrustes analysis [3]. We evaluate our method using a user study, comparing automatically and manually selected frames.

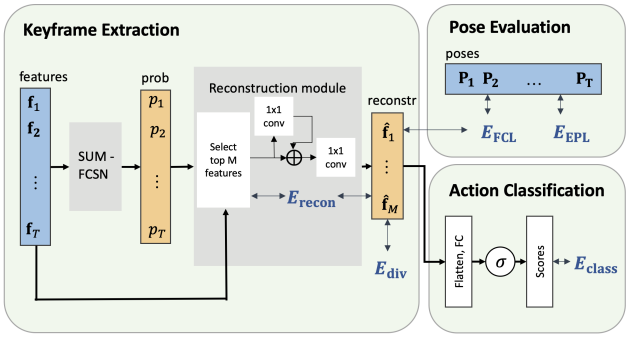

Figure 2: **System architecture.** The model contains three modules, keyframe extraction, pose evaluation, and action classification. The input are image features, $\mathbf{f}_1, ..., \mathbf{f}_T$, and pose vectors, $\mathbf{P}_1, ..., \mathbf{P}_T$, shown in blue boxes.

2 Methodology

Our model architecture consists of three modules: (i) a keyframe extraction module, (ii) a pose evaluation module, and (iii) a classification module. The keyframe extraction module applies the unsupervised version of fully convolutional sequence networks SUM-FCSN [18] and outputs a number of representative frames. The pose evaluation module computes posed-specific loss functions to re-weight the selected frames. The selected frames are passed through a classification module, which ensures that the frames are recognizable as the particular action.

As shown in Fig. 2, image features and pose coordinates, shown in blue boxes, are provided as input to the model. To extract image features we leverage a pretrained Temporal Segment Network (TSN) [22], which has shown high accuracy for action recognition. Following the common strategy for action recognition, we divide videos into T equal sized video clips and select a single frame from each clip. Thus, T frames are passed through the TSN to obtain feature vectors. The pose vectors for each of the T frames are obtained using AlphaPose [4]. Sections 2.1–2.3 detail the three modules of our method.

2.1 Keyframe extraction module

Our keyframe extraction module is built on the unsupervised version of a fully convolutional sequence network SUM-FCSN [18]. This network has an encoder-decoder architecture with 8 convolution layers in the encoder and 2 layers in the decoder. The model takes features $\mathbf{f}_1, ..., \mathbf{f}_T$ from T frames as input, estimates probabilities $p_1, ..., p_T$ for each frame, and reconstructs the image features, $\hat{\mathbf{f}}_1, ..., \hat{\mathbf{f}}_T$. The model as shown in Fig. 2, provides a probability value for each frame. The unsupervised version of the model (SUM-FCSN$_{unsup}$) does not require any keyframe labels. As defined in [18], two loss terms, reconstruction loss (E_{recon}) and diversity loss (E_{div}), ensure that the selected keyframes are both representative and diverse.

2.2 Pose Evaluation

To select suitable poses we introduce two loss terms, *Extreme Pose Loss* (E_{EPL}) and *Frame Contrastive Loss* (E_{FCL}). The Extreme Pose Loss is defined as the pose most distant to the mean pose. The 2D joint positions are extracted using AlphaPose [4, 24]. We map the set of pose coordinates $\{\mathbf{P}_t\}_{t=1}^{T}$ for T frames into shape space by applying centering, scaling and rotation. We first center the coordinates and scale them to unit length, obtaining vectors, $\mathbf{Z}_t = \frac{\mathbf{C}\mathbf{P}_t}{||\mathbf{C}\mathbf{P}_t||_2}$, where $\mathbf{C}$ is a centering matrix. Note that, $\mathbf{P}_t, \mathbf{Z}_t$ are of dimensions $K \times 2$ and $\mathbf{C} \in \mathbb{R}^{K \times K}$, where K is the number of joints in the pose. To map these vectors into shape space, each $\mathbf{Z}_t$ is aligned by applying an optimal rotation $\mathbf{\Gamma}_t \in SO(2)$. The set of optimal rotation matrices $\{\mathbf{\Gamma}_t\}$ for all pose coordinates is found by

$$\{\mathbf{\Gamma}_t\} = \min_{\mathbf{\Gamma}_t \in SO(2)} \sum_{t=1}^{T} \sum_{j=t+1}^{T} \Big[||\mathbf{Z}_t\mathbf{\Gamma}_t - \mathbf{Z}_j\mathbf{\Gamma}_j||_2 \Big]^2. \quad (1)$$

The Procrustes distance is defined as the distance in the shape space [3, 8]. The Fréchet mean, $\mathbf{P}^F$, defines the mean pose in the shape space, and our Extreme Pose Loss measures the Procrustes distance from this mean:

$$E_{\text{EPL}} = \frac{1}{T} \sum_{t=1}^{T} \exp\Big(- \frac{||\mathbf{Z}_t\mathbf{\Gamma}_t - \mathbf{P}^F||_2}{2\sigma^2} \Big), \quad (2)$$

where the Fréchet mean pose $\mathbf{P}^F$ is computed over the T frames, $\mathbf{P}^F = \frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_t\mathbf{\Gamma}_t$, and σ is empirically set to 10 by parameter grid search.

The second loss function is the Frame Contrastive Loss, E_{FCL}, which encourages low intra-action class difference of poses and image features, and high inter-action class difference. The Frame Contrastive Loss is defined as:

$$E_{\text{FCL}} = \frac{1 + \sum_{i,j \in \mathcal{V}} \mathbb{1}_{\{y_i = y_j\}} ||\mathbf{F}_i - \mathbf{F}_j||}{N_{\mathcal{V}}(1 + \sum_{i,j \in \mathcal{V}} \mathbb{1}_{\{y_i \neq y_j\}} ||\mathbf{F}_i - \mathbf{F}_j||)}, \quad (3)$$

where the features $\mathbf{F}_i = [\hat{\mathbf{f}}_t^i, \mathbf{P}_t^i]$ include both the reconstructed frame features, $\hat{\mathbf{f}}_t^i$, and the pose coordinates, $\mathbf{P}_t^i$, of the t^{th} frame in the i^{th} video. $\mathcal{V}$ denotes a set of indexes of $N_{\mathcal{V}}$ videos in a mini-batch, y_i is the label of the ith video in the mini-batch. The loss function minimizes the L2 distance of features in the same class and maximizes the distance of features belonging to different classes. To handle the case of a single element minibatch, we add 1 to both numerator and denominator. Pose estimation in video frames can be noisy and is affected by blur and occlusions. To mitigate

Figure 3: **Selected iconic poses.** (left) frames selected by using only reconstruction and diversity loss terms, E_{recon} and E_{div}, (right) frames selected by including, Extreme Pose Loss E_{EPL} and Frame Contrastive Loss E_{FCL}. Including these terms leads to selecting subjectively better frames with respect to pose and view.

these issues, we use pose coordinates with 80% of the individual points that have a confidence value above a threshold (0.7) and interpolate pose coordinates in frames with lower confidence. In videos where body pose detection fails for most frames, the Frame Contrastive Loss reverts to using only the frame features.

2.3 Classification module

The extraction of keyframes is weakly supervised by the action class labels. Reconstructed features of the top M frames selected by the keyframe extraction module and iconic pose selection module are passed through a classification module. This module passes the frames through a fully connected layer to predict the action class using a binary cross entropy loss function, E_{class}. The parameters of the keyframe extraction and action classification modules are updated during training. The complete model is trained end-to-end using the following loss function:

$$E_{\text{icon}} = \lambda_1 E_{\text{class}} + \lambda_2 E_{\text{div}} + \lambda_3 E_{\text{recon}} + \lambda_4 E_{\text{EPL}} + \lambda_5 E_{\text{FCL}} \quad (4)$$

where, $\lambda_1, \lambda_4 = 0.35$ and $\lambda_2, \lambda_3, \lambda_5 = 0.1$, found by hyperparameter search. At inference time, the frame features and pose coordinates are passed to the model, which predicts the probability p_i of each frame along with the classification label. The top M frames, ranked by probability, are selected. The frame with highest probability denotes the iconic pose frame.

3 Results

We carried out experiments on two public action recognition datasets, UCF-101 [21] and HMDB-51 [10]. The UCF-101 dataset contains 13,320 videos from 101 action classes. HMDB-51 [10] contains 6,849 clips from 51 action classes. For evaluating the performance of iconic pose frame selection, we conducted a user study, comparing automatically and manually selected frames.

Implementation details. We use pre-trained AlphaPose models [4] and TSN models [22] to obtain pose and image features respectively. The optimizer used for training the model is Adam [9] with learning rate 10^{-3} and $\beta_1 = 0.9$, $\beta_2 = 0.999$. The model was trained for 100 and 200 epochs for UCF101 and HMDB51 datasets, respectively. Training the model for 200 epochs takes approximately 16h on an RTX 2080Ti GPU.

3.1 Iconic pose selection

Figure 3 shows the frames selected by our model with two different loss functions from four videos of different action classes. The left image of each pair in Fig. 3 shows the iconic pose extracted by our model with $E_{\text{recon}}, E_{\text{div}}$ losses only. The right image of each pair shows the iconic poses extracted when including the pose loss terms E_{EPL} and E_{FCL}. Including these two additional terms results in selecting frames with better poses.

User Study. We carried out a preference study to evaluate our method relative to manually selected frames by annotators. We used 48 videos, containing 12 action classes, for which full body poses are in view, namely *LongJump, HighJump, PullUps, Dive, SwingBaseball, FlicFlac, Cartwheel, Somersault, Bowling, CleanAndJerk, GolfSwing* and *JavelinThrow*. For each video three annotators independently labeled the iconic pose frame manually. Annotators were given the instruction to select the frame that best represents the action shown in the video. We therefore obtained up to three different frames for each video, which we consider ground truth iconic pose frames. In 73% of cases (35 out of 48 videos) the proposed method obtained the same frame as that of one of the annotators. For the preference study we only use videos where our method selects different frames from manually annotated ones. In the experiment we sample pairs of images from the same video, selected by different methods: human-annotated, and two variations of our model, Iconic-RD and Iconic-RDEF. Iconic-RD denotes the model with only the reconstruction and diversity loss, Iconic-RDEF denotes the model with all loss terms. A total of 20 users were shown these image pairs and asked to select between the two frames. For each user, we randomly select 10 examples for each of the three pairwise comparisons, resulting in a total of 600 preference tests. The results in Table 1 show that the model including pose loss (Iconic-RDEF) improves over the model without pose loss (Iconic-RD) by a large margin. While Iconic-RDEF performs slightly worse in direct comparison to manually selected results,

Figure 4: **User study examples.** For each pair: (left) manually selected frames by annotators, (right) model selected frames. Frames favored by more users are shown with green border.

Table 1: **Iconic pose user study.** Each row shows the preference vote count for manually selected and model selected iconic pose frames. Iconic-RD and Iconic-RDEF are the models without and with pose loss terms, respectively.

	Manual	Iconic-RD	Iconic-RDEF
	126	74	
	104		96
		54	**146**
Preference	38.3%	20.5%	40.3%

it received the most votes overall.

Fig. 4 shows frames from the user study. The first two columns show cases where users preferred the model-selected frame over the manually annotated frame. The second two columns show cases where the majority of users preferred the manually annotated frames. In all examples iconic poses are extreme poses and similar iconic poses are obtained for the same classes. Example results on broadcast video of Olympic events are shown in Fig. 5. Frames are shown in temporal order and those selected by our model are shown in with a red border.

Effect of frame selection on action classification. We evaluate how the frame selection affects the action recognition accuracy in an additional experiment. As the selected number of frames, M, varies, the performance of action classification changes and we select M based on the highest accuracy. The performance is evaluated with respect to a Temporal Segmentation Network (TSN) model as baseline [22]. While TSN uses two streams for action recognition, spatial and temporal, we evaluate using the spatial stream only. Following the protocol in [19, 22], we first sample 25 frames from a video, each with 10 cropped versions for data augmentation. In TSN [22], the final model class is estimated by averaging the 25 scores. Our method selects frames from these 25 groups and estimates the performance from the selected set of scores.

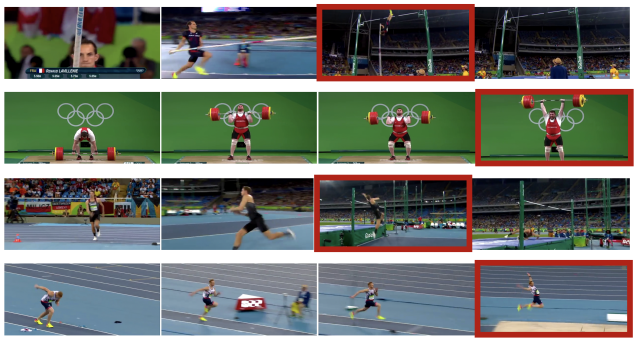

Figure 5: **Iconic pose frame selection results** (red border) on broadcast footage of Olympic events pole vault, weight lifting, high jump, and long jump.

Table 2: **Action classification accuracy.** The performance of our method when varying the number of selected frames, M, evaluated on UCF101/HMDB51 datasets, respectively.

	Split1	**Split2**	**Split3**	**Average**
TSN	85.5/54.4	84.9/50.0	84.5/49.2	85.1/51.0
M=25	86.9/54.5	83.3/50.3	84.7/50.2	85.0/51.7
M=20	87.8/54.9	83.5/50.5	85.1/52.0	85.5/52.5
M=16	87.9/55.0	84.2/50.8	**85.5/52.3**	85.9/52.7
M=12	**88.4/55.4**	**84.5/51.1**	85.0/51.9	**86.0/52.8**
M=8	87.3/53.7	83.5/50.5	85.0/51.0	85.1/51.7
M=4	87.2/52.8	83.5/50.6	83.8/50.5	84.8/51.3

Table 2 shows the recognition accuracy for different numbers of selected frames ($m = 4, 8, 12, 16, 20, 25$) on all the splits of UCF101 and HMDB51 datasets. The TSN row shows the performance of the baseline model. We report the performance of the spatial stream for TSN. For both datasets, the model with 12 selected frames performs the best on average.

Limitations: The extreme pose loss depends on accurate pose detection. For low quality videos, this process is noisy. For actions involving more than one person, such as *CricketShot, BasketballDunk, FrisbeeCatch*, we select the pose of the main actor based on size in the image. Generalizing to actions involving multiple actors is a possible extension of this work.

4 Conclusion

In this paper we introduced a method to automatically select an iconic pose frame from an action video. We introduced an Extreme Pose Loss term based on shape analysis of the pose coordinates and a Frame Contrastive Loss. We validate our design choices experimentally and demonstrate in a preference study that the method is comparable to human-selected frames. We have also shown experimentally that frame selection improves action recognition performance.

References

[1] Jackie Assa, Yaron Caspi, and Daniel Cohen-Or. Action synopsis: pose selection and illustration. *ACM Transactions on Graphics (TOG)*, 24(3):667–676, 2005.

[2] Shayok Chakraborty, Omesh Tickoo, and Ravi Iyer. Adaptive keyframe selection for video summarization. In *WACV*, pages 702–709, 2015.

[3] Ian L Dryden. *Statistical shape analysis*, volume 4. John Wiley & Sons, 1998. ISBN:978-0-471-95816-1.

[4] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[5] Yuli Gao, Tong Zhang, and Jun Xiao. Thematic video thumbnail selection. In *ICIP)*, pages 4333–4336, 2009.

[6] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *NeurIPS*, pages 2069–2077, 2014.

[7] Genliang Guan, Zhiyong Wang, Shiyang Lu, Jeremiah Da Deng, and David Dagan Feng. Keypoint-based keyframe selection. *IEEE TCSVT*, 23(4):729–734, 2012.

[8] Geethu Jacob and Sukhendu Das. Moving object segmentation in jittery videos by stabilizing trajectories modeled in kendall's shape space. In *BMVC*, 2017.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[10] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[11] Chunxi Liu, Qingming Huang, and Shuqiang Jiang. Query sensitive dynamic web video thumbnail generation. In *ICIP*, pages 2449–2452, 2011.

[12] David Liu, Gang Hua, and Tsuhan Chen. A hierarchical visual model for video object summarization. *IEEE TPAMI*, 32(12):2178–2190, 2010.

[13] Tianming Liu, Hong-Jiang Zhang, and Feihu Qi. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE TCSVT*, 13(10):1006–1013, 2003.

[14] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, pages 3707–3715, 2015.

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *CVPR*, pages 202–211, 2017.

[17] Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. From keyframes to key objects: Video summarization by representative object proposal selection. In *CVPR*, pages 1039–1048, 2016.

[18] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, pages 347–363, 2018.

[19] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

[20] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proc. 25th ACM Int. Conf. Information and Knowledge Management*, pages 659–668, 2016.

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.

[22] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[23] Wayne Wolf. Key frame selection by motion analysis. In *ICASSP*, volume 2, pages 1228–1231, 1996.

[24] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[25] Xiang Yan, Syed Zulqarnain Gilani, Hanlin Qin, Mingtao Feng, Liang Zhang, and Ajmal Mian. Deep keyframe detection in human action videos. *arXiv:1804.10021*, 2018.