Breaking the chain: liberation from the temporal Markov assumption for tracking human poses

Ryan Tokola
Oak Ridge National Laboratory
Oak Ridge, TN
tokolara@ornl.gov

Wongun Choi
NEC Research Labs
Cupertino, CA
wongun@nec-labs.com

Silvio Savarese
Dept. of Computer Science
Stanford University
Stanford, CA
ssilvio@stanford.edu

Abstract

We present an approach to multi-target tracking that has expressive potential beyond the capabilities of chain-shaped hidden Markov models, yet has significantly reduced complexity. Our framework, which we call tracking-by-selection, is similar to tracking-by-detection in that it separates the tasks of detection and tracking, but it shifts temporal reasoning from the tracking stage to the detection stage. The core feature of tracking-by-selection is that it reasons about path hypotheses that traverse the entire video instead of a chain of single-frame object hypotheses. A traditional chain-shaped tracking-by-detection model is only able to promote consistency between one frame and the next. In tracking-by-selection, path hypotheses exist across time, and encouraging long-term temporal consistency is as simple as rewarding path hypotheses with consistent image features. One additional advantage of tracking-by-selection is that it results in a dramatically simplified model that can be solved exactly. We adapt an existing tracking-by-detection model to the tracking-by-selection framework, and show improved performance on a challenging dataset (introduced in [18]).

1. Introduction

Tracking humans in video has been the focus of much recent attention in computer vision and robotics research, and its successful implementation has far-reaching areas of impact, such as security and surveillance, entertainment, video annotation and indexing, medical diagnosis, disability assistance, sociology and kinesiology, and autonomous navigation. Ideally, we would like tracking methods to work with unconstrained real-world video sequences with cluttered backgrounds and a moving camera. Unfortunately, tracking people is difficult because people move with complex dynamics, they have unpredictable appearances, and they are highly deformable.

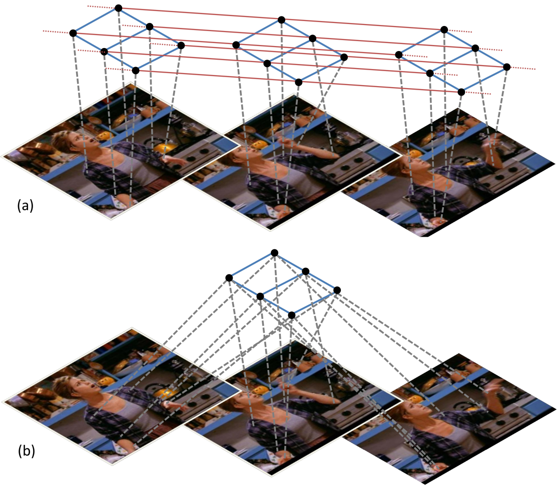

Figure 1. (a) The tracking-by-detection model from [18]. Each node represents the location of a body joint in one time frame. The nodes for each body joint form a chain in the time dimension (these edges are shown in red), following the temporal Markov assumption. This assumption means that consistency can only be enforced from one frame to the next. The spatial connections between joints (shown in blue) result in a model that is extremely loopy and must be solved approximately. (b) A tracking-by-selection reformulation of the model. Each node represents the path of a body joint through the video. With this model, the location of the body joint at any time does not depend upon its temporal neighbors. Instead, the chain for each joint has been broken and compressed into a single node. With this model, it is possible to reason about the entire trajectory of a joint. For example, trajectory hypotheses with consistent colors throughout the video can be rewarded. The model is much simpler, and can be solved exactly.

1.1. Tracking-by-detection

Recent approaches to tracking in video are dominated by tracking-by-detection, such as in [2], which separates the tasks of detection and tracking into two separate processes. Detections are made in each time frame, and then detections are associated to make coherent trajectories. The typical model for tracking-by-detection has the form of a chain-shaped hidden Markov model: the state at time t depends entirely upon observations at time t and the temporally adjacent states at times $t-1$ and $t+1$. The temporal Markovian assumption simplifies the model, and allows for efficient inference.

Chain-shaped tracking-by-detection models work well when detections are reliable, but they are limited by their simplifying assumptions — even though tracking is a fundamentally time-based operation, time is only used to enforce some degree of local consistency from one frame to the next. For example, a standard model will discourage large color changes between adjacent frames, but it is incapable of preventing a final trajectory that slowly changes from one color to a very different color. A more robust model could avoid this problem by adding dependencies. Unfortunately, simply adding this high order connectivity to traditional models would most often result in intractable inference.

1.2. Tracking-by-selection

We present a new approach to tracking that can enforce global (long term) consistencies across an entire video, yet is even simpler than a standard chain-shaped tracking-by-detection model. Recall that a tracking-by-detection framework will generate detection hypotheses in each frame and then create a chain of associations between them. In contrast, our approach, which we call *tracking-by-selection*, generates a set of *path hypotheses* instead of detection hypotheses. Each path hypothesis is a possible trajectory through time, whereas a detection hypothesis is a possible location at a specific time. The critical aspect of tracking-by-selection is that it breaks the chain of variables in traditional tracking approaches and replaces it with a single variable (see Figure 1). Information from the first frame is no longer linked only to the second frame — with path hypotheses an observation in the first frame can be freely compared to observations all along the trajectory. This means that we can easily reward globally consistent trajectories without the need for high order potentials.

Inference in the tracking-by-selection framework amounts to simply selecting a path hypothesis. In the case of a single-target tracking problem, inference is trivial: select the most likely path hypothesis for the target. In a multi-target tracking application, the model only requires reasoning about the relationships between targets (or body parts, in the case of tracking human poses), so inference involves selecting one path for each object. There is no need for the added complexity of time-based associations. The differences between tracking-by-detection and tracking-by-selection are illustrated in Figure 1.

Tracking-by-selection brings two significant contributions to the problem of tracking human poses:

Global consistency. Potentials that are unary and pairwise in our model can only be expressed as higher-order potentials in other models. These potentials can enforce consistency in ways that have previously been impossible. Because tracking-by-selection is grounded in paths through time and not static locations, each path hypothesis has access to information throughout the entire video.

The only way for a tracking-by-detection framework to enforce consistency across time is to build a model of the object and use the model as a prior for the remainder of the tracking procedure. These dynamic priors are frequently unreliable in the case of highly deformable objects, such as people.

Exact inference. Many models for tracking human poses that follow the tracking-by-detection framework must be solved with approximate inference algorithms, but the same models can be solved exactly if they are adapted to a tracking-by-selection framework. Because the model has been "collapsed" in time, inference in tracking-by-selection is dramatically simplified. The resulting model has the same structure as a single frame of the equivalent model in a tracking-by-detection framework. It may have cycles, but the only pairwise potentials to consider are those between objects.

The most obvious drawback to approximate inference is that there is no guarantee that the maximum a posteriori (MAP) solution will be found, but we also show in section 4 that some approaches to approximate inference can limit the performance of a model in more subtle ways.

2. Previous works

2.1. Human pose estimation in still images

Most contemporary pose estimators are indebted to the pictorial structures model [6], which introduced an efficient and effective way of modeling the relationships between body parts. Image parsing techniques in [14] and enhanced in [7] and [5] led to significant improvements in very challenging datasets. [16] shows impressive results by efficiently handling a rich set of image features. [25] abandons the traditional articulated limb model, and instead models the co-occurrence of oriented parts. [21] proposes a hierarchical Articulated Part-based Model for jointly detecting people and estimating their poses. [22] introduces an efficient branch-and-bound algorithm that can find exact solutions to complex models that can only be solved approximately with traditional inference methods.

2.2. Tracking human poses

[10] is one of the first works to attempt a model that enforces both spatial and temporal coherence between human body parts. It performs inference in an approximate manner by alternating between spatial and temporal optimizations.

Recent works often build dynamical models of specific activities, particularly those related to the walk cycle. [3] [8] [24] learn strong dynamical priors that provide assistance to the pose estimate when the image evidence is lacking. They are very good at tracking people in challenging videos, but they cannot cope with novel actions and generally do not adapt to style. [8] learns the correspondence between silhouette and 3D pose, as well as walk cycle. A stereo camera is used for the segmentation, and a skeleton is fit to the silhouette with the help of the dynamic priors. Many works such as [1] track human poses with an array of cameras. [19] integrates imagery and 3D range data to obtain accurate 3D pose localization.

[4] creates a more complex model by reasoning about the relationships between the poses of multiple people at the same time. This follows our reasoning that increased connectivity between observations can improve tracking, but they only add to the spatial connectivity of typical tracking models. The model is still fundamentally Markovian, so the enhanced spatial reasoning is unable to strongly propagate through time. [17] extend their cascaded pictorial structures framework to pose tracking by adding edges between body parts in time, and further refine their approach in [18]. Even though states in each frame are only connected to adjacent frames, the emphasis in [18] is on approximate inference methods, which are required for their sophisticated model. They decompose the graph into an ensemble of loopless tree structures, which are combined in different ways.

[13] shares some similarities with this paper, but each path that it generates for a body part is made by finding the shortest path through a set of detections. The path has a standard Markov chain structure, in that it is only capable of enforcing consistency among adjacent frames. With our work, however, we are able to select from among a multitude of hypotheses not only a single path that is consistent over the entire video, but also pairs of paths that have consistent image features between them. Furthermore, instead of fixing paths one or two at a time, our inference jointly optimizes the selection of paths for all body parts.

All of the above methods rely upon the temporal Markov assumption, and are only capable of enforcing appearance consistency between adjacent time frames. [15] is a rare example of a human pose tracker that uses an instance-specific model of the appearance of individual body parts. The appearance models are constructed by repeated iterations of detection and association. Like [18], their model is decomposed into a set of trees, and inference in approximate.

2.3. Tracking with tracklets

Many previous works, such as [12] have grouped detections into tracklets, which are then combined to form trajectories. In most cases, the same Markovian assumptions apply. [26] is capable of enforcing global appearance consistency across many frames, but the model is so complex that the trajectories must be built one target at a time, and there are no pairwise dependencies between targets.

3. The tracking-by-selection model

Tracking-by-detection divides the tracking problem into two stages. In the first, object hypotheses are generated in a frame, which dramatically reduces the size of the space that must be searched in the second stage, in which associations are built between hypotheses to create smooth tracks. We propose a new framework called *tracking-by-selection* that is similar in many respects, but shifts the burden of reasoning about time from the second association stage to the first hypothesis generation stage.

The two most important aspects of tracking-by-selection are: (a) global appearance consistency can be enforced without the need for high order potentials, and (b) tracking-by-selection uses a model that is much less complex than tracking-by-selection, which makes exact inference feasible.

In a tracking-by-detection framework, the goal of inference in tracking human poses is to pick one location for each body part in each time frame in such a way that the location of each body part is consistent with the image evidence (unary potential), locations of different body parts are consistent with each other (spatial pairwise potential), and the locations of one body part are consistent from one frame to the next (temporal pairwise potential). When framed as an energy maximization problem, a general tracking-by-detection model can be written as $\max_s E(s)$, where

$$E(s) = \sum_t \Big[\sum_p \Big(\theta_p^{a\top} \Psi(s_p^t) + \sum_q \theta_{pq}^{a\top} \Psi(s_p^t, s_q^t) + \theta_{pp}^{\top} \Psi(s_p^t, s_p^{t+1}) \Big) \Big], \tag{1}$$

where $\Psi(s_p^t)$ is the unary potential for body part p at time t, $\Psi(s_p^t, s_q^t)$ is the spatial pairwise potential between parts p and q at time t, and $\Psi(s_p^t, s_p^{t+1})$ is the temporal pairwise potential between the states of part p at time t and $t+1$. All θ are model parameters. In all equations we omit explicit references to image evidence for clarity.

The tracking-by-selection framework can be written as

$$E(s) = \sum_p \Big(\theta_p^{\top} \Psi(x_p) + \sum_q \theta_{pq}^{\top} \Psi(x_p, x_q) \Big), \tag{2}$$

where $\Psi(x_p)$ and $\Psi(x_p, x_q)$ are the unary and pairwise potentials for path hypotheses for body part p.

The potentials for an existing tracking-by-detection model can easily be incorporated into a tracking-by-selection framework. We decompose the potentials in Equation (2) by

$$\theta_p^\top \Psi(x_p) = \theta_p^{a\top} \Psi^a(x_p) + \theta_p^{b\top} \Psi^b(x_p) \tag{3}$$

and

$$\theta_{pq}^\top \Psi(x_p, x_q) = \theta_{pq}^{a\top} \Psi^a(x_p, x_q) + \theta_{pq}^{b\top} \Psi^b(x_p, x_q), \tag{4}$$

where

$$\Psi^a(x_p) = \frac{1}{T} \sum_t (\Psi(s_p^t)) \tag{5}$$

and

$$\Psi^a(x_p, x_q) = \frac{1}{T} \sum_t (\Psi(s_p^t, s_q^t)) \tag{6}$$

are the average unary and pairwise potential across the the path hypothesis. Note that the same model model parameters $\theta_p^{a\top}$ and $\theta_{pq}^{a\top}$ can be used for both tracking-by-detection and tracking-by-selection. $\Psi^b(x_p)$ and $\Psi^b(x_p, x_q)$ represent the potentials that are only available to a tracking-by-selection model. They will be discussed in more detail in sections 3.1 and 3.2

One important characteristic of the tracking-by-selection framework is that the final selection step has significantly lower complexity than tracking-by-detection. In effect, it has "collapsed" the entire model in the time dimension. The information that was contained in the temporal edges of a tracking-by-detection model can now be expressed as part of the unary potential of a path hypothesis, and all temporal edges are eliminated.

If a tracking-by-detection model, such as the graphical model in Figure 1(a), has six body parts, seven spatial pairwise potentials, and the video has 50 frames, then there is a total of $7 \cdot 50 + 6(50 - 1) = 644$ edges. The tracking-by-selection reformulation of the same model, shown in Figure 1(b), has only seven edges.

3.1. Unary global consistency

Tracking-by-selection is capable of expressing global image consistency with a single unary potential. For example, the unary potential for a path hypothesis may include the variance of image features along the extent of the path:

$$\Psi^b(x_p) = -\mathrm{var}(\mathbf{x}_p^1, ..., \mathbf{x}_p^T), \tag{7}$$

where $\mathbf{x}_p^t$ is an image feature at the path hypothesis of body part p at time t. In practice this potential is most useful for hands. It is unnecessary for shoulders, which easily lend themselves to consistent tracks, and it is occasionally damaging to tracks for elbows, which may have significant changes in appearance depending on the location and orientation of the forearm.

Naively rewarding a low color variance would most strongly support path hypotheses that stay "locked" to a part of the background, this is particularly true in the case of hands, which will naturally have some appearance variation as they move about. To remedy this, we reduce the reward for paths with little motion. Basing the motion estimate on total displacement would reward "jittery" paths, so instead we use the maximum displacement of a path over the course of a video.

$$\Psi^b(s_p) = -\frac{\mathrm{var}(\mathbf{x}_p^1, ..., \mathbf{x}_p^T)}{\max\limits_{t^1, t^2} \parallel x_p^{t1} - x_p^{t2} \parallel_2^2}, \tag{8}$$

A traditional tracking-by-detection model for human poses is necessarily cyclical and quite complicated, yet it is still unable to replicate these consistency and displacement features without the use of high order potentials, which would have the form $\Psi(s_p^1, s_p^2, ..., s_p^T)$ and would be extremely difficult to solve.

3.2. Pairwise global consistency

While unary potentials in a tracking-by-selection framework are capable of reasoning about joints over time, pairwise potentials can reason about the space *between* joints in time. For example, if a person has a blue shoulder, a purple elbow, and a yellow armband, we would like to enforce that shoulder and elbow are always blue and purple, and that a point between them is always yellow. To accomplish this, we consider the variance of each of N equally-spaced points between path hypotheses:

$$\Psi^b(x_p, x_q) = -\frac{\sum_{n=1}^N \mathrm{var}(\mathbf{x}_{(p,q)^1}^n, ..., \mathbf{x}_{(p,q)^T}^n)}{\max\limits_{t^1, t^2} \parallel x_q^{t1} - x_q^{t2} \parallel_2^2}, \tag{9}$$

As with to our unary potential for joint color consistency, we include motion as a feature. For this potential, we use the maximum displacement of the path of the "child" joint. This means that the motion of the elbow will modify the potential for the upper arm, and the motion of the hand will modify the potential for the lower arm.

Attempting to incorporate this potential into a chain-shaped model would require a potential of order $2T$: $\Psi(s_{p^1}^1, ..., s_{p^T}^T, s_{p^1}^1, ..., s_{p^T}^T)$.

3.3. Generating path hypotheses

The successful generation of plausible path hypotheses is clearly of paramount importance. It must be possible to generate many path hypotheses quickly, and they must be sufficiently diverse. Otherwise, it is possible that no path hypotheses may exist close to the best solution. We promote a sampling method, but it is important to note that this is an engineering decision, and may different approaches are possible.

Instead of only building path hypotheses from the first frame forward, we initialize path hypotheses in every frame. This helps ensure that some path hypotheses will pass through every image region with a sufficiently strong detector response.

In each frame, we normalize the single-frame unary potential for each joint into a probability and draw samples from it. Each sample represents the initialization of a new path hypothesis, which is propagated forward and backward in time. The propagation is driven by samples taken from a transition probability.

A transition probability based only on distance in the image plane may be sufficient for some videos with a static camera, but when the camera moves, rotates, or zooms the potential may result in a track that fails to follow the camera motion. Furthermore, even if the camera is still, there are time-based image features that can improve our estimate of where a point at time t will be at time $t + 1$. We model the transition probability by combining four observations: **Spatial distance**: the simple distance in image space. **Color distance**: the magnitude of the difference in color between the two points. **Spatial distance with optical flow**: the distance in image space after accounting for optical flow. **Shared video segments**: do two points both reside in the same video segment?

Video segments are essentially superpixels that exist in time. We use the video segments from [9]. Video segments are more robust to certain kinds of motion than optical flow.

The relative weights of these four observations were learned through the Structured Support Vector Machine (SSVM) framework [23]. The weights were learned using the ground truth locations of joints from a training set as examples, and the loss function was based on the Euclidean distance from ground truth. Separate sets of weights were learned for each joint type (shoulder, elbow, and wrist).

3.4. Inference

For inference in a full tracking-by-detection framework to be tractable it is common to use approximate inference, such as loopy belief propagation [20] or decomposition methods [18]. The equivalent tracking-by-selection model, on the other hand, has only as many pairwise potentials as appear in a single frame of the tracking-by-detection model. The remaining "figure eight" model shown in Figure 1(b) can be exactly solved with the junction tree algorithm, but the $\mathcal{O}(n^3)$ complexity of a straightforward application requires excessive memory when using a large number of path hypotheses. To keep the memory overhead manageable, the model is instead decomposed into an ensemble of trees. Similar to [11], the decomposition is accomplished by fixing one of the nodes representing the path of an elbow, as shown in Figure 2. One tree is generated for every possible state (path hypothesis) for the elbow. After the elbow node has been fixed (and therefore becomes an observation instead of a latent variable), the remainder of the model is a tree, and can therefore be efficiently solved with max-sum belief propagation. The max marginals resulting from inference on each tree can be directly compared, and the maximum of the max marginals is the solution. It is important to understand that unlike some other forms of decomposition-based inference, this procedure results in an exact solution. Inference over a model with 2,000 path hypotheses per joint takes only a couple of seconds in MATLAB when GPU-optimized functions are used.

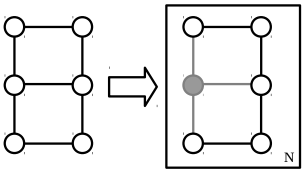

Figure 2. Exact inference on a tracking-by-selection model. By fixing the value of an elbow node, the 'figure eight" model becomes a tree that can efficiently be solved with max-sum belief propagation. The plate in the figure represents the N trees that are generated from each of the N path hypotheses for the elbow joint. Message passing is performed once for every possible state of the fixed elbow node, and taking the maximum of the max marginals results in an exact solution.

The structure of our particular problem lends itself to a particularly efficient inference method, but even if a model had additional cycles, it could still likely be solved exactly. Our specific inference method is not unimportant, but instead we wish to emphasize that with tracking-by-selection the final inference procedure is only as complex as a single frame of the tracking-by-detection equivalent.

3.5. Scalability

Clearly, a straightforward application of tracking-by-selection cannot be used with extremely long videos because a prohibitive number of path hypotheses would be required. In future work we will demonstrate that a sliding window-based approach can be used to apply tracking-by-selection to videos of arbitrary length.

4. Implementation

To demonstrate the advantages of tracking-by-selection, we adapt an existing tracking-by-detection model to the tracking-by-selection framework. This allows for a more direct and fair comparison than an entirely new model.

We use the publicly available code from [18], and replace the single-frame joint hypotheses with path hypotheses to create a tracking-by-selection model.

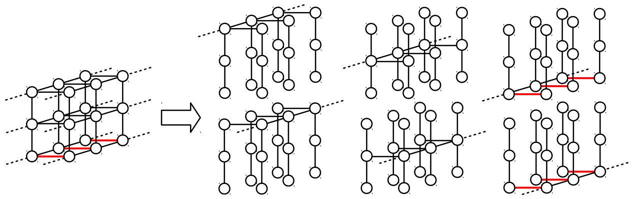

Figure 3. The overall model from [18] is decomposed into six submodels to allow for efficient inference. Note that each submodel contains a subset of the edges in the original. The red lines indicate the symmetrical potential between wrists, which only appear in two of the six submodels.

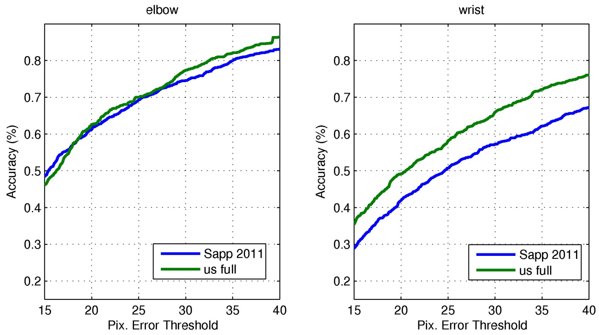

Figure 4. Main results: us vs. baseline [18]

It is important to note that the inference in [18] does not search over the dense pixel grid, but also restricts their search to a limited number of plausible single-frame hypotheses. We take this logic forward one step, and restrict our search to a set of multi-frame path hypotheses.

[18] shows substantial evidence to support the claim that approximate inference can perform nearly as well as exact inference in some cases. A small loss in performance is a small price to pay for substantially reduced computational burden. [18] decomposes their model into an ensemble of tree-shaped submodels, performs inference on each of the submodels, and constructs a final solution by enforcing consistency between the submodels. When dual decomposition is used the final inference is exact, but excessively slow, and long-running inference sessions were terminated before convergence. With other types of agreement, the solution is only approximate, but the performance is only slightly reduced.

One important aspect of [18] is that the submodels were trained independently of each other. By training the submodels in isolation, the model is not being optimized for exact inference. To illustrate this point, we focus on one component of the unary potential for wrists. [18] very correctly reason that optical flow can be a strong cue for detecting hands. In each frame, the gradient of the optical flow is filtered by a hand-shaped template. The magnitude of the response is used as a feature in the unary potential for wrists. This flow-based cue is helpful, but it is very sporadic. When a hand is moving rapidly it may have a very strong response, but without motion it gives no response at all. Because people have two hands, this variability can easily result in "double counting" if one hand is moving and the other is not — the flow-based feature will lead the pose estimate to locate both hands at the site of movement. There are two easy ways to counteract the double counting problem. The first is to only lightly weight the flow-based feature, and the other is to have a symmetrical repulsion potential between the two hands.

A repulsion potential, which penalizes two wrist estimates that are close to each other, is ideal for preventing double counting, but it only exists in two of the six submodels in [18], as shown in Figure 3. Four of the submodels, lacking the repulsion potential for wrists, will naturally learn a very low weight for the flow-based feature. Even if dual decomposition is used to force agreement between the submodels, the overall contribution of the flow-based feature will be quite weak. On the other hand, if training was performed using the entire model the flow-based feature would always be balanced by the symmetrical repulsion potential, and a more substantial weight could be learned.

To demonstrate the advantage of learning on a complete model, we re-learned parameters for the flow-based wrist feature and the symmetrical wrist potential using exact inference on our model. This resulted in greater weights for both and ultimately to improved performance. See Figure 7 for results comparing our learned parameters and parameters directly taken from [18].

Aside from the weights for the flow-based wrist feature and the symmetrical, model parameters were taken directly from [18]. Each submodel in [18] has separate learned parameters, so we averaged the parameters over all of the submodels.

Because only a few model parameters needed to be learned (the unary and pairwise consistency potentials from section 3.1 and 3.2, the flow-based wrist feature, and the symmetrical wrist potential), training was conducted with a simple grid search over possible parameter values. The objective function that was minimized was the total error in the training set.

5. Experiments

5.1. VideoPose2.0 dataset

Experiments were conducted on a variation of the VideoPose2.0 dataset introduced in [18]. The original dataset only includes alternating frames from the source videos, which is clearly not ideal for a system that is grounded in time-based features. The alternate version of the dataset, which is provided by the authors, includes the missing frames. In addition, the alternate version of the dataset con-

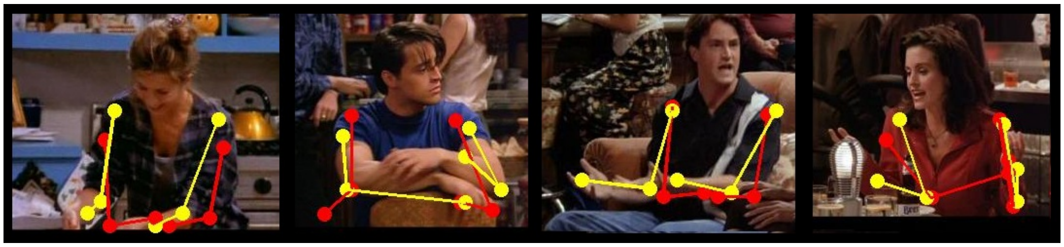

Figure 5. Anecdotal results. Our results are shown in yellow. Results from [18] are shown in red.

catenates several videos that were partitioned into smaller clips in the original version of the dataset. For a fair comparison to [18], their publicly available code was evaluated on the alternate version of the dataset, and it is these results that are shown here. The public code from [18] does not implement all of the methods found in the paper, but the results for the best-performing methods were all very similar. It is in fact the purpose of the paper to show that the simpler approximate methods perform as well as other more sophisticated methods.

5.2. Main results

Fig. 4 shows our main results. For these results, 1000 path hypotheses were used for each shoulder, and 2000 path hypotheses were used for each elbow and wrist. The figures plot the percentage of joints (the y-axis) that were correctly located within a certain number of pixels (the x-axis) of the ground truth. Clearly, the most important thing to note is the significant improvement in hand tracking. It can be seen that that the performance for elbows is similar to the baseline. In the VideoPose2.0 dataset, the motion of the elbows is much less significant than that of the hands, so enforcing spatio-temporal consistency will have less effect.

Some anecdotal results are shown in Fig. 5. Note that the image on the far right is a partial failure case. One of the main drawbacks of strongly enforcing global image consistency is that occlusions can lead to failure. Remember, however, that the generation of path hypotheses is stochastic. This means that one of the many path hypotheses may not be significantly affected by a brief occlusion.

5.3. Testing path hypotheses

Reducing the search space in a tracking problem to a relatively small set of paths is potentially troubling. After all, the solution in a tracking-by-selection framework can only be as good as the path hypotheses. To asses the quality of path hypotheses and estimate the performance ceiling of tracking-by-selection, Figure 6 shows experimental results for paths that have been selected by an oracle. Results are shown for groups of varying size. It can be seen that even with only 100 paths per body joint to choose from, a solution exists that is very close to the ground truth. It is interesting to note that increasing the number of paths provides a more dramatic boost to wrists than to elbows. There are two primary reasons for this. First, there is much greater spatial variation in the wrist locations, as the actors tend to gesticulate wildly. Second, wrists are simply harder to detect, since they tend to move rapidly and are one more joint removed from the relatively stable shoulders.

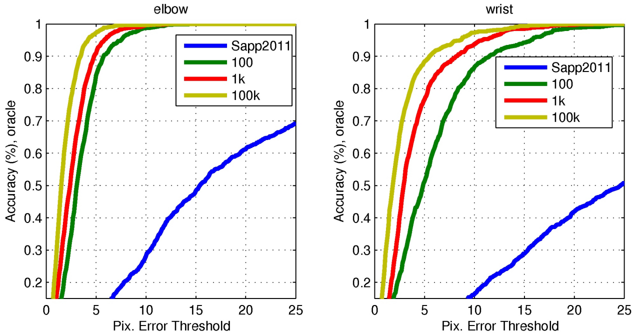

Figure 6. Results showing performance with an oracle selecting the best path with a varying number of random path hypotheses. The results from [18] are shown for reference.

5.4. System analysis

Figure 7 shows results for our model with one component removed at at time. The results for wrists and elbows are averaged. "new pairwise" refers to the pairwise potential shown in Equation 9. "new unary" refers to the pairwise potential shown in Equation 8. "extra hand weights" refers to the parameters that were re-learned with exact inference as discussed in section 4.

6. Conclusions

We have presented *tracking-by-selection*, a new framework for tracking human poses in video. By reasoning about path hypotheses instead of single-frame state hypotheses, racking-by-selection is capable of enforcing global consistency in ways that are intractable in traditional chain-shaped tracking-by-detection frameworks. We show that converting a human pose tracking system to a tracking-by-selection model results in improved performance on challenging tracking problems.

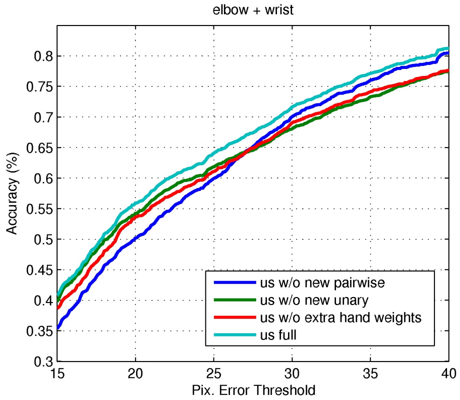

Figure 7. Results from our full system and results with individual features removed.

Acknowledgments

We acknowledge the support of NSF CAREER grant #1054127, NSF CPS grant #0931474 and a KLA-Tencor Fellowship.

References

[1] S. Allin, N. Baker, E. Eckel, and D. Ramanan. Robust tracking of the upper limb for functional stroke assessment. In *IEEE Transactions on Neural Systems and Rehabilitation Engineering (NSRE)*, 2010.

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, 2008.

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1014–1021, 2009. ID: 1.

[4] M. Andriluka and L. Sigal. Human context: Modeling human-human interactions for monocular 3d pose estimation. In *VII Conference on Articulated Motion and Deformable Objects (AMDO)*, Mallorca, Spain, July/2012 2012. Springer-Verlag, Berlin Heidelberg, Springer-Verlag, Berlin Heidelberg. Best paper award.

[5] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *Proc. BMVC*, pages 3.1–3.11, 2009. doi:10.5244/C.23.3.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[8] S. Gammeter, K. Schindler, L. V. Gool, A. Ess, B. Leibe, and T. Jaggli. Articulated multi-body tracking under egomotion. In *International Conference on Computer Vision (ICCV), 2008*, 2008.

[9] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. *IEEE CVPR*, 2010.

[10] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 690–695. IEEE, 2001.

[11] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 470–477. IEEE, 2005.

[12] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[13] V. Ramakrishna, T. Kanade, and Y. Sheikh. Tracking human pose by tracking symmetric parts. June 2013.

[14] D. Ramanan. Learning to parse images of articulated bodies. In *In NIPS 2007*. NIPS, 2006.

[15] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–467–II–474 vol.2, 2003. ID: 1.

[16] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[17] B. Sapp, D. Weiss, and B. Taskar. Sidestepping intractable inference with structured ensemble cascades. In *NIPS*, 2010.

[18] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[19] J. Shotton and T. Sharp. Real-time human pose recognition in parts from single depth images. *Training*, 2:1297–1304, 2011.

[20] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–421–I–428 Vol.1, 2004. ID: 1.

[21] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.

[22] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *CVPR*, 2012.

[23] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 104–, New York, NY, USA, 2004. ACM.

[24] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 238–245, 2006. ID: 1.

[25] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[26] A. R. Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: global multi-object tracking using generalized minimum clique graphs. In *Computer Vision–ECCV 2012*, pages 343–356. Springer, 2012.