Bilinear Programming for Human Activity Recognition with Unknown MRF Graphs

Zhenhua Wang, Qinfeng Shi, Chunhua Shen and Anton van den Hengel
School of Computer Science, The University of Adelaide, Australia
{zhenhua.wang01,javen.shi,chunhua.shen,anton.vandenhengel}@adelaide.edu.au

Abstract

Markov Random Fields (MRFs) have been successfully applied to human activity modelling, largely due to their ability to model complex dependencies and deal with local uncertainty. However, the underlying graph structure is often manually specified, or automatically constructed by heuristics. We show, instead, that learning an MRF graph and performing MAP inference can be achieved simultaneously by solving a bilinear program. Equipped with the bilinear program based MAP inference for an unknown graph, we show how to estimate parameters efficiently and effectively with a latent structural SVM. We apply our techniques to predict sport moves (such as serve, volley in tennis) and human activity in TV episodes (such as kiss, hug and Hi-Five). Experimental results show the proposed method outperforms the state-of-the-art.

1. Introduction

Human activity recognition (HAR) is an important part of many applications such as video surveillance, key event detection, patient monitoring systems, transportation control, and scene understanding. HAR involves several subfields in computer vision and pattern recognition, including feature extraction and representation, human body detection and tracking, and has attracted much research attention in recent years. A suite of methods have been proposed for this task and promising recognition rates have been achieved. For example, in [11], more than 91% of actions can be correctly classified for the KTH dataset [15]. A detailed review of different methods for HAR can be found in [1]. In this work, we focus on recognizing individual activities in videos which contain multiple persons who may interact with each other.

If one assumes that the activity of each person is an independent and identically distributed (i.i.d.) random variable from an unknown but fixed underlying distribution, one can perform activity recognition by training a multiclass classifier, or a group of binary classifiers, based on descriptors such as HoG, HoF [6] or STIP [10] extracted from body centred areas, as in [15], [12], [11]. We label these methods as i.i.d. as they classify the action of each person separately. These methods are thus very efficient, but they may not be effective as the i.i.d. assumption is often not true in practice. For example in a badminton game, a player smashing suggests that his opponent is more likely to perform bending and lobbing instead of serving.

Unlike traditional i.i.d methods, Markov Random Fields (MRFs) [8] are able to model complex dependencies and deal with local uncertainty in a consistent and principled manner. A possible MRF is shown in Figure 1 (d), where nodes represent the activity random variables, and the edges reflect the dependencies. The reason that MRFs can achieve a superior result to i.i.d. methods is because MRFs obtain the joint optimum over all variables whereas the i.i.d.
methods only seek optima for each variable separately. In
[5], MRFs are used to model activities like walking, queue-
ing, crossing etc. The MRFs' graphs are built simply based
on heuristics such as two person's relative position. These
heuristics may not be always reliable. For example, persons
waving to each other can be far away from each other. One
may deploy other heuristics, but they typically require spe-
cific domain knowledge, and apply only to a narrow range
of activities in a particular environment. Instead of using
heuristics, Lan et al. try to learn graphs based on the poten-
tial functions of the MRFs [9]. They seek the graph and the
activity labels that give the highest overall potential function
value (i.e. the smallest energy). The joint optimisation is
very different from a typical inference problem where only
labels are to be predicted. Thus they solve the joint opti-
misation problem approximately using a coordinate ascent
method by holding the graph fixed and updating the labels,
then holding labels fixed and updating the graph. Though
predicting each person's activities is only a by-product —
predicting group activities is their task — their idea is gen-
erally applicable to HAR. A key problem with this method
is that coordinate ascent only finds local optima. Since the
graphs of MRFs encode the dependencies, constructing re-
liable graphs is crucial. A MRF with an incorrect graph
(such as Figure 1 (b) (c) ) may produce inferior results to
an i.i.d. method ( in Figure 1 (a) ), whereas a MRF with a
correct graph ( in Figure 1 (d) ) produces a superior result, at
least for this example.

In this paper, we show Maximum A Posteriori (MAP) in-
ference for the activity labels and estimating the MRF graph
structure can be carried out simultaneously as a joint opti-
misation and that achieving the global optimum is guaran-
teed. We formulate the joint optimisation problem as a
bilinear program. Our bilinear formulation is inspired by recent work in Linear Program (LP) relaxation based
MAP inference with known graphs in [7] and [16], and be-
ief propagation (BP) based MAP inference with unknown
graphs [9]. We show how to solve the bilinear program effi-
ciently via the branch and bound algorithm which is guaran-
teed to achieve a global optimum. We then apply this novel
inference technique to HAR with an unknown graph. Our
experimental results on synthetic and real data, including
tennis, badminton and TV episodes, show the capability of
our method.

The reminder of the paper is organized as follows. We
first introduce our task and the model representation in Sec-
ton 2.1 followed by the MAP problem and its LP relaxation
in Section 2.2. Then we give our bilinear formulation for
the MAP inference with unknown graphs in Section 3.1. In
Section 3.2 we show how to relax the bilinear program to
an LP. In Section 3.3 we describe how to solve the bilinear
program. In Section 4 we show how to train the model (i.e.
parameter estimation) for the MRFs. Section 5 provides the
experimental results followed by conclusions in Section 6.

2. Modelling HAR with unknown MRF graph

2.1. The Model

In HAR our goal is to estimate the activities of m per-
sons \( Y = (y_1, y_2, \ldots, y_m) \in Y \), given an observation im-
age \( X \in X \). To model the dependencies between activities,
we use a MRF with the graph \( G = (\mathcal{V}, \mathcal{E}) \) where the ver-
tex set \( \mathcal{V} = \{1, 2, \ldots, m\} \) and the edge set \( \mathcal{E} \) is yet to be
determined. We cast the estimation problem as finding a
discriminative function \( F(X, Y; G) \) such that for an image
\( X \), we assign the activities \( Y \) and the graph \( G \) that exhibit
the best score w.r.t. \( F \),

\[
(Y^*, G^*) = \arg \max_{Y \in \mathcal{Y}, G \in \mathcal{G}} F(X, Y, G).
\]  

(1)

As in many learning methods, we consider functions linear
in some feature representation \( \Psi \),

\[
F(X, Y, G; \mathbf{w}) = \mathbf{w}^\top \Psi(X, Y, G).
\]  

(2)

Here we consider the feature map \( \Psi(X, Y, G) \) below,

\[
\Psi(X, Y, G) = \left[ \sum_{i \in \mathcal{V}} \Psi_1(X, y_i); \sum_{(i, j) \in \mathcal{E}} \Psi_2(X, y_i, y_j) \right].
\]  

(3)

The discriminative function \( F \) can be expressed as

\[
F(X, Y, G; \mathbf{w}) = \sum_{i \in \mathcal{V}} \mathbf{w}_1^\top \Psi_1(X, y_i) + \sum_{(i, j) \in \mathcal{E}} \mathbf{w}_2^\top \Psi_2(X, y_i, y_j),
\]  

(4)

where \( \mathbf{w} = [\mathbf{w}_1; \mathbf{w}_2] \). Now, (1) is equivalent to

\[
(Y^*, G^*) = \arg \min_{Y, G} \sum_{(i, j) \in \mathcal{E}} E_{ij}(y_i, y_j) + \sum_{i \in \mathcal{V}} E_i(y_i),
\]  

(5)

which becomes an energy minimisation problem with un-
known graph \( G \). Here \( E_{ij}(y_i, y_j) \) is the energy func-
tion over edge \((i, j) \in \mathcal{E}\) and the \( E_i(y_i) \) is the node energy
function over vertex \( i \in \mathcal{V} \).

2.2. The MAP inference and its LP Relaxation

If \( G \) is known, the MAP problem becomes

\[
Y^* = \arg \min_Y \sum_{(i, j) \in \mathcal{E}} E_{ij}(y_i, y_j) + \sum_{i \in \mathcal{V}} E_i(y_i).
\]  

(6)
A typical LP relaxation [16] of the MAP problem is
\[
\min_q \sum_{(i,j) \in E} q_{i,j}(y_i, y_j) E_{i,j}(y_i, y_j) + \\
\sum_{i \in V} \sum_{y_i} q_i(y_i) E_i(y_i)
\]
\[\text{s.t. } q_{i,j}(y_i, y_j) \in [0, 1], \sum_{y_i} q_{i,j}(y_i, y_j) = 1, \quad (9b)\]
\[\sum_{i} \sum_{y_i} q_i(y_i) E_i(y_i) \geq \sum_{i} \sum_{y_i} z_{i,j} \in [0, 1], \forall (i, j) \in E, \forall (i, j) \in E.\]
\[\sum_{i} \sum_{y_i} q_i(y_i) E_i(y_i) \geq \sum_{i} \sum_{y_i} z_{i,j} \in [0, 1], \forall (i, j) \in E, \forall (i, j) \in E.\]
This problem is a BLP with disjoint constraints i.e., constraints on \(q\) do not involve \(q\) and vice versa.

Solving the BLP in (9) returns \(\{q_i(y_i) \mid \forall i \in V, y_i \in \{0, 1\}\}\), \(\{q_{i,j}(y_i, y_j) \mid \forall i, j \in V, y_i, y_j \in \{0, 1\}\}\), and \(\{z_{i,j} \mid \forall i, j \in V\}\). We now show how to obtain the graph \(G^*\) and the MAP \(Y^*\) (i.e., best activities) from BLP outcomes.

Obtaining the graph We start with \(E^* = \emptyset\), \(\forall i, j \in V, i \neq j, z_{i,j} \geq 0.5\), \(E^* = E^* \cup \{(i, j)\}\). Thus we have the estimated graph \(G^* = (V, E^*)\).

Obtaining the MAP Assume \(y_i \in \{1, 2, \cdots, K\}\), then \(\forall i \in V, y_i^* = \arg\max_{y_i} \sum_{k=1}^{m} q_k(y_i)\). We have the estimated label \(Y^* = (y_1^*, y_2^*, \cdots, y_m^*)\).

3. Bilinear reformulation and LP relaxation

In this section we show how the MAP inference problem with an unknown graph can be formulated as a bilinear program (BLP), which can be further relaxed to an LP. We will also show how to obtain the labels and graph from the solution of the BLP.

3.1. Bilinear program reformulation

Before we solve (5), let us consider a simpler case first where the graph is unknown but the MAP solution \(Y^*\) is known. We first introduce variables \(\{z_{i,j}\}_{i,j}\) indicating the edge \((i,j)\) exists or not, for all \(i, j \in V\). Then the graph \(G\) can be found by seeking the set of edges that give the smallest energy as an integer program:
\[
\min_z \sum_{i \in V} \sum_{j \in V} z_{i,j} E_{i,j}(y_i^*, y_j^*)
\]
\[\text{s.t. } z_{i,j} \leq d, z_{i,j} = z_{j,i}, z_{i,j} \in \{0, 1\}, \forall i, j \in V, y_i^*, y_j^*.\]
Here we set \(E_{i,j}(y_i, y_j) = +\infty\) when \(i = j\) since \((i, i)\) is not an edge. \(E_i(y_i)\) can be ignored, since it is independent of the choice of edges. \(d \in \mathbb{N}\) is a preset number that enforces the maximum degree of a vertex, allowing one to enforce a sparse structure. However dropping the degree constraint does not change the nature of the problem. This integer program can be relaxed to a linear program by changing the domain of \(z_{i,j}\) from \(\{0, 1\}\) to \([0, 1]\).

From (7) and (8), we can see that (5) can be relaxed to the problem below,
\[
\min f(q, z) = \sum_{i,j \in V} q_{i,j}(y_i, y_j) E_{i,j}(y_i, y_j) z_{i,j}
\]
\[+ \sum_{i \in V} \sum_{y_i} q_i(y_i) E_i(y_i).\]
\[
\begin{align*}
\min_{q,z} f(q, z) \leq f(q^*, z^*) & \quad \text{since } (q^*, z^*) \in \Omega. \quad \text{(11)}
\end{align*}
\]

Since (11) is a relaxation of (9), \( \Phi_{\text{lb}}(\Omega) \leq \min_{q,z \in \Omega} f(q, z) \). Thus condition (12a) is satisfied. By the argument ofLemma 1 in [4], Condition (12b) is also satisfied. Hence the convergence holds.

### 4. Training

We now present a maximum margin training method for predicting structured output variables, such as human activity labels. Given \( \ell \) image-activities pairs, \( \{(X^i, Y^i)\}_{i=1}^\ell \) (note that the graphs are not known), we estimate \( w \) via Latent Structural Support Vector Machine (LSSVM) [19].

\[
\begin{align*}
\min_{w} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{\ell} \left[ \max_{Y,G} \Psi(X^i, Y, G) + \Delta(Y^i, Y) \right], \quad (13)
\end{align*}
\]

where \([a]_+ = \max\{0, a\}\). Here \( C \) is the trade-off between the regulariser and the risk, and \( \Delta \) is the label cost that penalises incorrect activity labels. Here we use a well known hamming distance for the label cost, that is

\[
\begin{align*}
\Delta(Y^i, Y) = \frac{1}{m} \sum_{j=1}^{m} \delta(y_j^i \neq y_j), \quad (14)
\end{align*}
\]

where the indicator \( \delta(\cdot) = 1 \) if the statement is true, 0 otherwise. We follow [19] in solving (13) via Convex-Concave Procedure (CCCP) [20] due to non-convexity of (13).

CCCP requires solving the following two problems:

\[
\begin{align*}
\max_{G} \Psi(X^i, Y^i, G). \quad (15)
\end{align*}
\]

\[
\max_{Y,G} \Psi(X^k, Y, G) + \Delta(Y^k, Y). \quad (16)
\]

Here (15) is efficiently solved via (8). Clearly (16) is equivalent to

\[
\begin{align*}
\min_{Y,G} \sum_{(i,j) \in E} E_{i,j}(y_i, y_j) \quad (17)
\end{align*}
\]

where \( E_{i,j}(y_i) = E_i(y_i) - \delta(y_j^i \neq y_j) \). Note that (17) has the same form as (5), and thus can also be formulated as a bilinear program which in turn can be solved by the Branch and Bound method.

**Lan’s method** The approach closest to ours is that proposed by Lan et al. in [9] using LSSVM via CCCP to recognise human activities. They also are required to solve (15) and (16). They use an LP similar to (8) to solve (15), and use a coordinate ascent style algorithm to approximately
solve (16). The coordinate ascent style algorithm iterates the following two steps: (1) holding $G^*$ fixed and solving,
\[
Y^* = \arg \max_Y W^T \Psi(X^k, Y, G) + \Delta (Y^k, Y)
\]
(18)
via belief propagation; (2) holding $Y^*$ fixed and solving
\[
G^* = \arg \max_G W^T \Psi(X^i, Y^*, G)
\]
via an LP (8). It is known that coordinate ascent style algorithms are prone to converging to local optima. In the experiment section we provide an empirical comparison between Lan’s method and ours.

5. Experiments

We apply our method to a synthetic dataset and three real datasets. The real datasets include two sports competition datasets (tennis and badminton), and a TV episode dataset.

5.1. Synthetic data

To quantitatively evaluate the performance of different methods for MAP inference and graph estimation, we randomly generate node (unary) and edge (binary) energies for different scales of problems. Specifically, 8 groups of energies are generated and each group corresponds to a fixed number of nodes and a fixed number of activities. For each group we randomly generate energies 50 times. The ground truth graphs and activities labels are obtained by exhaustive search for (5). For the true graph $G = (V, E)$, predicted graph $G' = (V, E')$ with $m$ nodes, true labels $Y$ and predicted label $Y'$, we define two errors below,
\[
e_y(G, G') = \frac{1}{m(m-1)} \sum_{i,j \in V} | z_{i,j} - z'_{i,j} |,
\]
(20)
\[
e_l(Y, Y') = \frac{1}{m} \sum_{i=1}^{m} \delta(y_i \neq y'_i),
\]
(21)
for which we call them the graph error and the label error respectively.

Here we compare four methods: our bilinear method (BLP), Lan’s method in [9] (Lan’s), the isolated graph (0 edge connections) with node status decided according to node potentials (we call it the Isolate method), belief propagation on fully connected graphs (BP+Full). We report the average errors (for 50 runs) for the 12 groups of synthetic data in Fig. 2. As expected, our bilinear method (the green curves) outperforms all other methods on both graph prediction and activity prediction. The superior results of our method are due to the use of global optimisation techniques, whereas Lan’s method achieves only local optima. The errors for both Isolate and BP+Full are high, because their graphs are not learnt. This shows the importance of learning the graph structure. An interesting observation is that the label prediction error of the Isolate method is much lower than that of BP+Full, though their performance on graph estimation is reversed. This may be caused by (loopy) belief propagation on graphs with many loops — fully connected graphs are used in BP+Full.

5.2. Predict moves in sports

The task is to find sporting activity labels for $m$ persons $Y = (y_1, y_2, \ldots, y_m) \in \mathcal{Y}$, given an observation image $X \in \mathcal{X}$. The bounding box for each athlete is already given.

5.2.1 Datasets

Two datasets are used here. The first one is the UIUC badminton match dataset [17]. We only use the annotated
men’s singles video with 3,072 frames. This dataset includes four moves: smash (SM), backhand (BH), forehand (FH) and others (OT). The second dataset is created by us. This dataset is a mixed-doubles tennis competition video of 2961 frames. We have annotated this video frame by frame. The annotation includes the position and the size of each body-centred bounding box, the pose and the action of each player. This dataset includes four moves: normal hit (HT), serve (SV), volley (VL) and others (OT). Thus the moves for the \( i \)-th person can be expressed as a discrete variable \( y_i \in \{1, 2, 3, 4\} \).

### 5.2.2 Features

We use the LSSVM (Section 4) to train a discriminative model to predict moves. The feature \( \Psi \) in (3) consists of two local features \( \Psi_1, \Psi_2 \) defined below,

\[
\Psi_1(X, y_i) = s_i \otimes e_1(y_i),
\]

\[
\Psi_2(X, y_i, y_j) = t_i \otimes t_j \otimes e_1(y_i) \otimes e_1(y_j) \otimes e_2(r_{i,j}),
\]

- **Move Local Appearance:** Similar to [9], \( s_i = (s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4}) \in \mathbb{R}^4 \), where \( s_{i,j}; (j = 1, 2, 3, 4) \) is a move confidence score of assigning the \( i \)-th person the \( j \)-th move. The score is the discriminative function value of a SVM classifier trained on the local image descriptor similar to that in [17]. The only difference is that we use histogram the dense gradients, rather than the silhouette of the human body area, since estimating silhouette is tricky when both camera and people are moving. All settings for the feature extraction process are as suggested in the literature. We pick this descriptor for two reasons. First, the descriptor is extracted from a stack of 15 consecutive frames, which accounts for the temporal action context. Second the length of descriptor vector is well controlled via the PCA projection.

- **Pose:** \( t_i \) is a vector of body pose confidence scores based on a SVM classifier trained on the same local descriptor used for \( s \). Here five discrete body poses are considered: profile-left, profile-right, frontal-left, frontal-right and backwards, thus \( t_i \in \mathbb{R}^5 \).

- **Relative position:** Here \( r_{i,j} \in \{1, 2, \ldots, 6\} \) is the relative position of person \( i \) and person \( j \). There are six possible relative positions including overlap, near-left, near-right, adjacent-left, adjacent-right and far. The relative position of two persons is determined according to their 2D Euclidean distance as in [14].

- **Tensor product:** Here \( e_1(y_i) \) is a 4-dimensional vector with 1 in the \( y_i \)-th dimension, and 0 elsewhere. Likewise, \( e_2(r_{i,j}) \) is a 6-dimensional vector with 1 in the \( r_{i,j} \)-th dimension, and 0 elsewhere. \( \otimes \) denotes the Kronecker tensor.

Intuitively, \( \Psi_1 \) reflects the confidence of assigning one person to different moves, \( \Psi_2 \) captures the co-occurrence of the related persons’ body poses, relative 2D position and their moves. Based on this joint feature representation, our model predict the moves via (5).

All datasets are randomly split into two parts: one for training and the other for testing. We compare four methods:

- **MCSVM** uses a multiclass SVM trained on the local image descriptor.
- **SSVM** uses a structural SVM [18] for training. Finding the most violated constraint requires inference which is done via BP. The graph is constructed by finding the minimum spanning tree weighted by the 2D Euclidean distance between persons.
- Lan’s method described in Section 4. The degree of the vertex \( d \) is set to 1.0. This means each node can have at most two edges ( \( z_{i,j} \geq 0.5 \) adds an edge).
- **BLP** (our method) described in Section 4. We use Mosek package [2] to solve LP involved in our bilinear method. We set the degree \( d \) in Eq. (11b) being 1.0.

The confusion matrices for two datasets are presented in Table 1 and 2. We can see that in tennis dataset, our BLP outperforms the other methods on all four moves. In badminton dataset, our BLP outperforms the other methods on 3 moves: FH, BH and SM. SSVM achieves the best recognition rate on OT, and our BLP achieves the second best. As expected, MCSVM performs the worst in general, since it relies on the local descriptors. An interesting observation is that Lan’s method does not always outperform SSVM.

### 5.3. Predict activities in TV episodes

The TVHI dataset proposed in [14] is a benchmark for predicting HAR in real TV episodes. It contains 300 short videos collected from TV episodes and includes five activities: handshake (HS), hug (HG), High-Five (HF), kiss (KS) and No-Interaction (NO). Each of the first four activities consists of 50 video clips. No-Interaction contains 100 video clips. We also bisect this dataset into training and testing parts as was done in the sports competition datasets. We use the same four methods as in the sports competition datasets. The confusion matrices are presented in Table 3. Our BLP outperforms the others on all activities. Lan’s method performs reasonably well on four activities: HS, HG, HF, and KS, but performs poorly on NO.

We show prediction results in Fig. 3 for four methods on different datasets. A green node means correct prediction and a red node means wrong prediction. We can
see that when the graph is learnt correctly, MRF-based approaches (see in the 2nd row, column 3, 4) outperform the i.i.d. method MCSVM (see in the 3rd column) than the graphs built from heuristics in SSVM (see the 2nd column). Overall, our BLP learns the most accurate graph and therefore makes the most accurate activity prediction.

6. Conclusion and future work

The structure of the graph used is critical to the success of any MRF-based approach to human activity recognition, because they encapsulate the relationships between the activities of multiple participants. These graphs are often manually specified, or automatically constructed by heuristics, but both approaches have their limitations. We have thus shown that it is possible to develop a MAP inference method for unknown graphs, and reformulated the problem of finding MAP solution and the best graph jointly as a bilinear program. Using the bilinear program based MAP inference, we have shown that it is possible to estimate parameters efficiently and effectively with a latent structural SVM. Applications in predicting sport moves and human activities in TV episodes have shown the strength of our method.

The BLP formulation is not only applicable to MRFs, but also to graphical models with factor graphs and directed graphs (i.e. Bayesian networks). One possible future work is to seek BP style algorithms to replace LP relaxations in solving BLP. Another future direction is to consider temporal dependencies among video frames using tracking techniques and dynamic hierarchy graphical models.

Acknowledgement This work was supported by Australian Research Council grants DP1094764, DE120101161, FT120100969, and DP11010352.

Table 1. Confusion matrices of the tennis dataset (sports match)

<table>
<thead>
<tr>
<th>Alg.</th>
<th>MCSVM</th>
<th>SSVM</th>
<th>Lan’s</th>
<th>BLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT</td>
<td>OT</td>
<td>OT</td>
<td>OT</td>
<td>OT</td>
</tr>
<tr>
<td>SV</td>
<td>SV</td>
<td>SV</td>
<td>SV</td>
<td>SV</td>
</tr>
<tr>
<td>HT</td>
<td>HT</td>
<td>HT</td>
<td>HT</td>
<td>HT</td>
</tr>
<tr>
<td>VL</td>
<td>VL</td>
<td>VL</td>
<td>VL</td>
<td>VL</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrices of the badminton dataset (sports match)

<table>
<thead>
<tr>
<th>Alg.</th>
<th>MCSVM</th>
<th>SSVM</th>
<th>Lan’s</th>
<th>BLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT</td>
<td>OT</td>
<td>OT</td>
<td>OT</td>
<td>OT</td>
</tr>
<tr>
<td>SV</td>
<td>SV</td>
<td>SV</td>
<td>SV</td>
<td>SV</td>
</tr>
<tr>
<td>HT</td>
<td>HT</td>
<td>HT</td>
<td>HT</td>
<td>HT</td>
</tr>
<tr>
<td>VL</td>
<td>VL</td>
<td>VL</td>
<td>VL</td>
<td>VL</td>
</tr>
</tbody>
</table>

References

Table 3. Confusion matrices of the TVHI dataset (television episodes)

<table>
<thead>
<tr>
<th>Alg</th>
<th>MCSVM</th>
<th>SSVM</th>
<th>Lan's</th>
<th>BLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/A</td>
<td>NO</td>
<td>HS</td>
<td>HF</td>
<td>HG</td>
</tr>
<tr>
<td>NO</td>
<td>0.37</td>
<td>0.07</td>
<td>0.21</td>
<td>0.11</td>
</tr>
<tr>
<td>HS</td>
<td>0.01</td>
<td>0.55</td>
<td>0.06</td>
<td>0.17</td>
</tr>
<tr>
<td>HF</td>
<td>0.09</td>
<td>0.03</td>
<td>0.52</td>
<td>0.21</td>
</tr>
<tr>
<td>HG</td>
<td>0.02</td>
<td>0.14</td>
<td>0.20</td>
<td>0.49</td>
</tr>
<tr>
<td>KS</td>
<td>0.07</td>
<td>0.11</td>
<td>0.09</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Figure 3. Prediction results of different methods on tennis data (1st two rows), badminton data (the 3rd row) and TV episode data (the last 2 rows). **First Column:** the i.i.d. method using MCSVM; **Second Column:** SSVM; **Third Column:** Lan’s method; **Last Column:** our BLP method. Both the nodes and edges of the MRFs are shown. For the i.i.d. method, there are no edges. The green node indicates correct predictions, and the red nodes incorrect predictions.


