Multimedia Event Detection using Visual Features
Dishan Gupta, Rahul Goutam and Amos Ng
Language Technologies Institute, Carnegie Mellon University

INTRODUCTION
What is MED
- Activity-centered happening involving people and process-driven action.
- Detect occurrence of event within a video clip (NIST).

Using unsupervised feature learning
- More generalizable to different domains.
- Less time-consuming.
- Less expensive.

Independent Subspace Analysis to learn features [4].

DATASET OVERVIEW

EVENT NAME: ATTENDING A BOARD TRICK (181)

Figure 2: An overview of the TRECVID 2011 MED dataset [6].

Our dataset consists of 3622 video clips from the NIST TRECVID 2011 MED task dataset spanning 15 distinct events and each clip belonging to only one of them. The events are shown in Figure 2.

UNSUPERVISED FEATURE EXTRACTION [1]

ISA is a 3-layer neural network [4], with square and square-root non-linearities in the second and third layers, respectively.

Objective Function: In detail, given an input pattern \( x \), the output layer activation is given by:

\begin{equation}
 p_i(x; W, V) = \sum_k W_i k x_j V_k x_j \end{equation}

The weights \( W \) are learned, whereas \( V \) are fixed in order to account for the subspace structure of neurons in the second layer. Thus, we are presented with the following optimization problem:

\begin{equation}
 \text{minimize}_{W, V} \sum_i p_i(x; W, V) \end{equation}

\begin{equation}
 \text{subject to } W V = I \end{equation}

The objective function (eq. (2)) is convex, but normal batch gradient descent would not be able to satisfy its constraint eq. (3). Thus, during each iteration we are required to update \( W \) by projecting it into the constraint set:

\begin{equation}
 W = s(W^V)^T W \end{equation}

Figure 3: A standard ISA network. The pooling units take input from two simple units to represent the subspace structure of 2.

CLASSIFICATION

- Features extracted using ISA algorithm used for classification.
- SVM with an exponential \( \gamma \) kernel.

RESULTS

Table 1: Accuracy and average precision results on five events.

<table>
<thead>
<tr>
<th>Event</th>
<th>Accuracy</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempting a Board Trick</td>
<td>77.40%</td>
<td>0.32</td>
</tr>
<tr>
<td>Feeding an Animal</td>
<td>77.40%</td>
<td>0.33</td>
</tr>
<tr>
<td>Landing a Fish</td>
<td>82.88%</td>
<td>0.26</td>
</tr>
<tr>
<td>Wedding Ceremony</td>
<td>82.20%</td>
<td>0.28</td>
</tr>
<tr>
<td>Working on a Woodworking</td>
<td>80.13%</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Figure 5: Precision (y-axis) vs Recall (x-axis) plots for four events.

REFERENCES