MAX-PRODUCT DYNAMICAL SYSTEMS AND APPLICATIONS TO AUDIO-VISUAL SALIENT EVENT DETECTION IN VIDEOS

Petros Maragos and Petros Koutras
School of ECE, National Technical University of Athens, 15773 Athens, Greece
maragos@cs.ntua.gr, pkoutras@cs.ntua.gr

ABSTRACT
This paper introduces a theory for max-product systems by analyzing them as discrete-time nonlinear dynamical systems that obey a superposition of a weighted maximum and evolve on nonlinear spaces which we call complete weighted lattices. Special cases of such systems have found applications in speech recognition as weighted finite-state transducers and in belief propagation on graphical models. Our theoretical approach establishes their representation in state and input-output spaces using monotone lattice operators, finds analytically their state and output responses using nonlinear convolutions, studies their stability, and provides optimal solutions to solving max-product matrix equations. Further, we apply these systems to extend the Viterbi algorithm in HMMs by adding control inputs and model cognitive processes such as detecting audio and visual salient events in multimodal video streams, which shows good performance as compared to human attention.

Index Terms— nonlinear systems, multimedia signal processing, lattices, minimax algebra, event detection, cognitive modeling.

1. INTRODUCTION AND SUMMARY
Several successful algorithms in pattern recognition and machine learning are based on a max-product arithmetic. Examples include speech recognition using weighted finite-state transducers (WFSTs) [32, 20], belief propagation in probabilistic graphical models [3, 40], and the maximum approximation used by the Viterbi decoding algorithm for likelihood scores during state estimation [33]. Further in signal processing and control there are several established areas using max/min superpositions and related operations of signals or vectors; examples include (i) the max-plus convolution (a.k.a. dilation) in morphological signal/image processing [18, 28, 36, 38] convex analysis [26, 35] and optimization [1], (ii) the minimax algebra used in scheduling [12], and (iii) the max-plus control in discrete-event dynamical systems [11, 23, 9]. Further, in multimodal signal processing for cognition modeling, which has been a main motivation for this work, several psychophysical and computational experiments indicate that the superposition of sensory signals or cognitive states seems to be better modeled using max or min rules, possibly weighted. Such an example is the recent work [15] on attention-based multimodal video summarization where a (possibly weighted) min/max fusion of features from the audio and visual signal channels and of salient events from various modalities seems to outperform linear fusion schemes. Finally, the sensory-semantic integration problem in multimedia signal processing requires fusion of two different continuous modalities (audio and vision) with discrete language symbols and semantics extracted from text. Similarly, in control and robotics there are efforts to develop hybrid systems that can model interactions between heterogeneous information streams like continuous inputs and symbolic strings [5]. In both of these applications we need models where the computations among modalities/states can handle both real numbers and Boolean variables; this is possible using max/min rules.

Motivated by the above multimodal signal processing problems, in this paper we develop some theoretical tools for the representation and analysis of nonlinear systems whose dynamics evolve based on the following state-space max-product model:

\[
x(t) = A(t) \boxplus x(t-1) \lor B(t) \boxtimes u(t)
\]
\[
y(t) = C(t) \boxtimes x(t) \lor D(t) \boxtimes u(t)
\]

(1)

where \( t \) denotes a discrete time index, \( \lor \) denotes maximum, \( x(t) \) is an evolving state vector, \( u(t) \) is the input signal (scalar or vector), \( y(t) \) is an output signal (scalar or vector), and \( A, B, C, D \) are appropriately sized matrices. \( \boxplus \) denotes the following nonlinear matrix product with max-product operations:

\[
P = Q \boxplus R, \quad p_{ij} = \bigvee_k q_{ik} \times r_{kj}
\]

(2)

The state equations (1) are written for the case of time-varying coefficients. If the matrices are constant and under zero-initial conditions, the input-output relationship of (1) can be described by a max-product convolution:

\[
y(t) = (h \otimes u)(t) = \bigvee_k u(k)h(t-k)
\]

(3)

where \( h \) is the system’s impulse response. By replacing maximum (\( \lor \)) with minimum (\( \land \)) in (1) and (3) we can also obtain a dual model that describes the dynamics of min-product systems.

Compare the above with linear systems [4, 6, 22, 17], which deal with linear maps: \( x(t) = Ax(t-1) + Bu(t) \) and \( y(t) = Cx(t) + Du(t) \). There, all the matrix-vector products and signal convolutions are linear, based on a sum-of-products arithmetic.

A max-product system is a special case of more general systems, studied in detail in [30], whose algebra is based on maximum of * operations. Examples of ‘multiplication’ * include the sum and the product, but * may be only a semigroup operation. The resulting algebras include the max-plus algebra \( \mathbb{R} \cup \{-\infty, \max, +\} \) used in scheduling and operations research [12], discrete-event dynamical systems [10, 11, 8, 9], automated manufacturing [23, 24, 13] and max-plus control [10, 16, 7]; the min-plus algebra or else known as tropical semiring \( \mathbb{R} \cup \{+\infty\}, \min, +\} \) used in shortest paths on networks [12] and in natural language processing [32, 20]; the fuzzy

This research work was supported by the project COGNIMUSE which is implemented under the ARISTEIA Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.
logic semiring \([\{0, 1\}, \lor, T]\) with statistical \(T\)-norms playing the role of fuzzy intersection used in fuzzy automata and neural nets [25, 21], and fuzzy dynamical systems [31].

**Our Contributions.** (1) Developed a theory for max-product systems analyzing both their dynamics in state-space and their input-output convolutional representation by using a new and powerful class of underlying spaces, the *complete weighted lattices* (CWLs). The detailed theory of CWLs is developed in [29, 30] to which we refer the reader for all proofs. (2) Derived analytic formulae for computing the state and output responses of max-product systems as well as for finding their input-output max-product convolutions, represented in both cases via lattice monotone operators in adjunction pairs. Further, use the latter to generate optimal solutions for solving max-product equations \(A \boxtimes x = b\). (3) Studied various control-theoretic issues of max-product systems. (4) Developed applications of max-product systems that extend the Viterbi algorithm of hidden Markov models (HMMs) to cases with control inputs and can estimate the saliences of audio-visual events in multimodal videos with good performance as compared to human attention.

2. BACKGROUND ON LATTICES AND OPERATORS

The background material in this section follows [2], [37], [19], [18] and [29]. A partially-ordered set, briefly poset \((P, \leq)\), is a set \(P\) in which a partial ordering \(\leq\) is defined. If the ordering \(\leq\) is total, then we have a chain. A lattice is a poset \((L, \leq)\) any two of whose elements have a *supremum* (a.k.a. least upper bound), denoted by \(X \lor Y\), and an *infimum* (a.k.a. greatest lower bound), denoted by \(X \land Y\). We often denote the lattice structure by \((L, \lor, \land)\). A lattice \(L\) is complete if each of its (finite or infinite) subsets has a supremum and an infimum in \(L\).

**Duality:** In any lattice \(L\), by replacing the partial ordering \(\leq\) with its dual \(\preceq\) and by interchanging the roles of the supremum and infimum, we can form a new lattice called the dual lattice and often denoted by \(L'\). To every definition, property and statement that applies to \(L\) there also corresponds a dual one that applies to \(L'\).

**Examples of Complete Lattices:** (a) The chain of extended real numbers \(\mathbb{R} = \mathbb{R} \cup \{\infty, -\infty\}\) equipped with the natural order \(\leq\). (b) The power set \(P(E) = \{X : X \subseteq E\}\) of an arbitrary set \(E\) equipped with the partial order of set inclusion where the supremum and infimum are the set union and intersection. 

**Function Lattices:** The set of discrete-time signals \(f : \mathbb{Z} \to \mathbb{R}\) equipped with the pointwise ordering \(\leq\), supremum and infimum of \(\mathbb{R}\).

**Increasing Operators:** Given two operators \(\psi\) and \(\phi\) on a complete lattice \(L\) we can define pointwise a partial ordering \(\leq\) between them, their supremum \((\psi \lor \phi)\) and infimum \((\psi \land \phi)\). Further, we define the composition of two operators as an operator product: \(\psi(\phi(X)) \equiv \psi(\phi(X))\); special cases are the operator powers \(\psi^n = \psi \circ \psi^{n-1}\). Some useful types and properties of lattice operators \(\psi\) include the following: (i) identity: \(1d(X) = X \forall X \in L\). (ii) extensive: \(\psi \geq 1d\). (iii) anti-extensive: \(\psi \leq 1d\). (iv) idempotent: \(\psi^2 = \psi\).

A lattice operator \(\psi\) is called increasing if it is order-preserving, i.e. \(X \leq Y \Rightarrow \psi(X) \leq \psi(Y)\). Four important types of increasing operators are the following:

\[
\begin{align*}
\delta & \text{ is dilation iff } \delta(V_i, X_i) = V_i, \delta(X_i) \\
\varepsilon & \text{ is erosion iff } \varepsilon(\Lambda_i, X_i) = \Lambda_i, \varepsilon(X_i) \\
\alpha & \text{ is opening iff } \alpha \text{ is increasing, idempotent & anti-extensive} \\
\beta & \text{ is closing iff } \beta \text{ is increasing, idempotent & extensive}
\end{align*}
\]

The four above types of lattice operators were originally defined in [37, 18] as generalizations of the corresponding standard morphological image operators.

Dilations and erosions come in pairs as the following concept reveals. The pair \((\varepsilon, \delta)\) of two operators \(\delta\) and \(\varepsilon\) on a complete lattice \(L\) is called an *adjunction* on \(L\) if

\[
\delta(X) \leq Y \iff X \leq \varepsilon(Y) \quad \forall X, Y \in L
\]

In any adjunction \((\varepsilon, \delta)\), \(\varepsilon\) is called the *adjoint erosion* of \(\delta\), whereas \(\delta\) is the adjoint dilation of \(\varepsilon\). There is a one-to-one correspondence between the two operators of an adjunction, since, given a dilation \(\delta\), there is a unique erosion

\[
\varepsilon(Y) = \bigvee \{X \in L : \delta(X) \leq Y\}
\]

such that \((\varepsilon, \delta)\) is adjunction, and vice-versa.

From the composition of the erosion and dilation of any adjunction \((\varepsilon, \delta)\) we can generate an opening \(\alpha = \delta \varepsilon\); since \(\alpha\) is an opening, we have \(\alpha(f) \leq f\) and \(\alpha = \alpha'\). Dually, any adjunction can also generate a closing \(\beta = \varepsilon \delta\). Both of these are special cases of morphological filters in [37, 18], a.k.a. *lattice projections* [29], since they are increasing and idempotent.

3. THEORY OF MAX-PRODUCT SYSTEMS

3.1. Weighted Lattices of Vectors and Signals

All elements of the vectors, matrices, or signals involved in the description of max-product systems take their values from the set \(\mathbb{K} = [0, \infty]\) of nonnegative extended reals. We equip \(\mathbb{K}\) with the following scalar operations: (A) the standard maximum or supremum \(\lor\) or \(\max\), which plays the role of a generalized ‘addition’; (A’) the standard minimum or infimum \(\land\) or \(\min\). It plays the role of a generalized ‘dual addition’. (M) the multiplication \(\times\) or \(\times\) extended over \([0, \infty]\) which has 1 as its identity and 0 as its null element, and distributes over any supremum. (M’) a ‘dual multiplication’ \(\times\) which has \(\infty\) as null element, distributes over any infimum and coincides with \(\times\) on \((0, \infty)\). The four above operations make \(\mathbb{K}\) an algebraic structure called *clodum* (complete lattice-ordered double monoid) [27, 29]. We can also define a conjugation operation mapping bijectively each element \(a\) to its conjugate element \(a = 1/a\) in \(\mathbb{K}\) equipped with the natural order \(\leq\). This interchanges suprema with infima; further \(a \times b = a^{-1} \times b^{-1}\). In \([0, \infty]\) the \(\times\) and \(\times\) operations coincide in all cases with only one exception, the multiplication of 0 with \(\infty\). Thus, henceforth we shall use only one multiplication \(\times\) and remember that the case \(0 \times \infty\) will have value 0 (resp. \(\infty\)) if it is combined with other terms via a supremum (resp. infimum).

Consider the set \(W\) consisting of all nonnegative functions \(F : E \to \mathbb{K}\) defined on an arbitrary nonempty set \(E\) and taking values in the clodum \(\mathbb{K} = [0, \infty]\). If we extend pointwise the supremum \((\lor \lor G)\) and infimum \((\land \land G)\) and scalar multiplication \((a \times F)\) for functions \(F, G \in W\) and scalars \(a \in \mathbb{K}\), the set \(W\) becomes a complete weighted lattice (CWL) over \(\mathbb{K}\). We can also have conjugation of functions by defining \(T(l) = 1/F(l)\). The axioms of CWLs bear a remarkable conceptual similarity with those of linear spaces as analyzed in our recent work [29, 30]. We focus on two special cases: (i) if \(E = \{1, 2, \ldots, n\}\), then \(W\) becomes the set of all \(n\)-dimensional vectors with elements from \(\mathbb{K}\). (ii) If \(E = \mathbb{Z}\), then \(W\) becomes the set of all discrete-time signals with values from \(\mathbb{K}\).

On linear spaces, a linear system \(\Gamma\) obeys linear superposition:

\[
\Gamma(\sum_i a_i F_i) = \sum_i a_i \Gamma(F_i)
\]

On a CWL the conceptually analogous superposition would be to have systems \(\delta\) that obey a max-product superposition:

\[
\delta(\bigvee_i c_i F_i) = \bigvee_i c_i \delta(F_i)
\]
This means that $\hat{\theta}$ is both a dilation and an inversion to vertical scalings (in short V-scalings) of signals $F(t) \mapsto aF(t)$. We call $\hat{\theta}$ a dilation V-scaling invariant (DVI) system.

**CWL of vectors:** Consider now the CWL vector space $W = K^n$, equipped with the pointwise partial ordering $x \leq y$, supremum $x \lor y = [x, y]$, and infimum $x \land y = [x, y]$, and scalar multiplications of vectors. On finite-dimensional linear vector spaces a vector map is linear iff it can be represented as a linear product between the system’s matrix and the input vector. Similarly, we have shown that on the CWL $W$ a map is DVI iff it can be represented as the max-product between the system’s matrix and the input vector.

**Theorem 1** (a) The vector $V$ is linear iff it can be represented as a linear product between the input vector $x$ and the matrix $M = [m_{ij}]$ with $m_{ij} = (\delta(u_j))$, where $u_j$ are basis vectors. This map is a vector dilation $\delta_M(x) = M \otimes x$. Its adjoint vector erosion, so that $(\varepsilon, \hat{\delta})$ is an adjunction, can be shown to equal [30]

$$
\varepsilon(y) = M^* \otimes' y, \quad M^* = \overline{M^T},
$$

(8)

where $M^* \triangleq [m^{-1}_{ji}]$ is the adjoint matrix of $M = [m_{ij}]$, and $\otimes'$ denotes the matrix min-product; namely, $P = Q \otimes' R$ with $p_{ij} = \wedge_k q_{ik} r_{kj}$. This adjunction helps us solve max-product equations:

$$
A \otimes x = b
$$

(9)

Often (9) does not have an exact solution, in which case we can find an optimum approximate solution by solving the following constrained minimization problem:

Minimize $\|A \otimes x - b\|

subject to $A \otimes x \leq b$

(10)

where $\| \cdot \|$ is either the $\ell_\infty$ or the $\ell_1$ norm.

**Theorem 2** (a) The vector $x = A^+ \otimes b$ is a solution to (10).

(b) If Eq. (9) has a solution, then $x$ is its greatest sub-solution.

Our method for solving (10) is to consider vectors $x$ that are sub-solutions in the sense that $A \otimes x \leq b$ and find the greatest such sub-solution using adjunctions. The set of sub-solutions forms a semigroup under vector $\lor$ whose supremum equals $x$, which yields either the greatest exact solution of (9) or an optimum approximate solution in the sense of (10). This adjunction-based solution creates a lattice projection via the opening $\hat{\delta}(\varepsilon(b)) \leq b$ that best approximates $b$ from below.

**CWL of signals:** Consider the set $W$ of all discrete-time signals $f : Z \to K$ with values from $K = [0, \infty)$. Equipped with pointwise supremum $\lor$ and pointwise scalar multiplications, this becomes a complete weighted lattice. The signal translations are the operators $\tau_h(f)(t) = f(t - k)$. A signal operator on $W$ is called translation invariant iff it commutes with any such translation. This translation-invariance contains both a vertical translation and a horizontal translation which is the well-known time-invariance. Now, if $g(t)$ is the impulse, equal to 1 at $t = 0$ and 0 elsewhere, every signal $f$ can be represented as a supremum of translated impulses

$$
f(t) = \bigvee_k f(k) g(t - k)
$$

(11)

Consider now operators $\Delta$ on $W$ that are dilations and translation-invariant in the above sense. Then, $\Delta$ is both DVI in the sense of (7) and time-invariant. We call such operators dilation translation-invariant (DTI) systems. Applying $\Delta$ to an input signal $f$ decomposed as in (11) yields the output as the max-product convolution $\otimes$ of the input with the system’s impulse response $h = \Delta(q)$:

$$
\Delta(f)(t) = (f \otimes h)(t) = \bigvee_k f(k) h(t - k)
$$

(12)

**Theorem 2** A signal operator $\Delta$ is a DTI system iff it can be represented as the max-product convolution of the input signal with the system’s impulse response $h = \Delta(q)$.

### 3.2. State and Output Responses

Based on the state-space model of a max-product dynamical system (1), we can compactly express its state response and output response if we know its transition matrix:

$$
\Phi(t_2, t_1) \triangleq \left\{ \begin{array}{ll} A(t_2) \otimes \cdots \otimes A(t_1 + 1) & \text{if } t_2 > t_1 \\ I_n & \text{if } t_2 = t_1 \end{array} \right. \quad (13)
$$

for $t_2 \geq t_1$, where $I_n$ is the $n \times n$ identity matrix. By using induction on (1), the state and output responses of the time-varying nonhomogeneous system can be found, for $t \geq 0$,

$$
x(t) = \Phi(t, 0) \otimes x(0) \lor \left( \bigvee_{i=1}^t \Phi(t, i) \otimes B(i) \otimes u(i) \right) \quad (14)
$$

$$
y(t) = C(t) \otimes \Phi(t, 0) \otimes x(0) \lor D(t) \otimes u(t) \lor \left( \bigvee_{i=1}^t C(t) \otimes \Phi(t, i) \otimes B(i) \otimes u(i) \right) \quad (15)
$$

The zero-state part of $y$ is a time-varying max-product convolution. If matrices $A, B, C, D$ are constant, the state equations become:

$$
x(t) = A \otimes x(t - 1) \lor B \otimes u(t)
$$

$$
y(t) = C \otimes x(t) \lor D \otimes u(t)
$$

(16)

and $\Phi(t_2, t_1) = A^{(t_2 - t_1)}$, where $A^{(t)}$ denotes the $t$-fold max-product of $A$ with itself. By representing the matrix-vector max-product as a dilation operator $x \mapsto \delta_A(x) = A \otimes x$, the solutions of the constant-matrix state equations become

$$
x(t) = \delta_A[x(0)] \lor \left( \bigvee_{i=1}^t \delta_A^{t-i} \delta_B[u(i)] \right) \quad (17)
$$

$$
y(t) = \delta_C \delta_A^{t} [x(0)] \lor \left( \bigvee_{i=1}^t \delta_C \delta_A^{t-i} \delta_B[u(i)] \right) \lor \delta_D[u(t)] \quad \text{zero-input resp.}
$$

Thus, the output response is found to consist of two parts: (i) the zero-input response which is due only to the initial conditions $x(0)$ and assumes a zero input, and (ii) the zero-state response which is due only to the input $u(t)$ and assumes zero initial conditions $x(0)$.

For single-input single-output systems the mapping $u(t) \mapsto y_{ss}(t)$ can be viewed as a translation invariant dilation system $\Delta$. Hence, the zero-state response can be found as the max-product convolution of the input with the system’s impulse response $h = \Delta(q)$. The latter can be found from the general output by setting initial conditions $x(0) = 0$ and the input $u(t) = q(t)$:

$$
h(t) = \left\{ \begin{array}{ll} D(t) & t = 0 \\ C \otimes A^{(t)} \otimes B, & t \geq 1 \end{array} \right. \quad (18)
$$

The previous results allowed us to address and solve in [30] various important control-theoretic problems for max-product systems, such as their stability, controllability and observability. We outline next the stability result. A useful bound for signals $f(t)$ processed by such systems is their supremal value $\bigvee f(t)$. We call max-product systems bounded-input bounded-output (BIBO) stable iff an upper bounded input yields an upper bounded output, i.e. if

$$
\bigvee u(t) < \infty \implies \bigvee y(t) < \infty
$$

(19)

Since all signals involved are nonnegative, the above definition of sup-stability coincides with their absolute stability.
4. HMMS EXTENSIONS AND APPLICATIONS TO DETECTING MULTIMODAL SALIENCIES

Assume a video sequence of audio-visual events each to be scored with some degree of saliency in [0,1] where ‘saliency’ is some bottom-up low-level sensory form of attention by a human watching this video. The states $x_1, x_2, x_3, x_4$ represent time-evolving monomodal or multi-modal saliencies, where 1=audio, 2=visual, 3=audiovisual, and 4=non-salient. Peaks in these saliency trajectories signify important events, which can be automatically detected and produce video summaries that agree well with human attention [15]. The following state equations are a possible time-varying max-product dynamical model we propose for the evolution of these saliencies:

$$x_i(t) = \left( \begin{array}{c} a_{i1}x_1(t-1) \\ a_{i2}x_2(t-1) \\ a_{i3}x_3(t-1) \\ a_{i4}x_4(t-1) \end{array} \right) \ast p_i(t) \vee \left( \begin{array}{c} b_{i1}u_1(t) \\ b_{i2}u_2(t) \\ b_{i3}u_3(t) \end{array} \right)$$

for state $i = 1, 2, 3, 4$. The constants $a_{ij}$ represent state transitions and $p_i(t)$ denotes the probability of state $x_i(t)$ being salient based on observed measurable low-level feature vectors $o_i$. We assume that the parameters $a_{ij}$ and $p_i(t)$ are given. The operation $\ast$ must distribute over $\vee$ and can be a product, min or max.

Assume first that $\ast$ is the product. Given a time sequence of observations $(o_1, o_2, \ldots, o_t)$ one can fit HMMS to these data using maximum likelihood [34]. Then, the first term in the RHS of (20) models the evolution of the Viterbi dynamic programming (DP) algorithm used in automatic speech recognition with HMMS for optimal state estimation, if we initialize at $t=0$ the four states by setting $x_i(0) = \pi_i p_i(0)$ where $\pi_i$ denotes the probability of the system being at the $i$th state at $t=0$. For example, if the inputs $u_i(t)$ are all zero, then the single output $y(t) = \bigvee_i x_i(t)$ computes the Viterbi score, which is the probability for having observed the data $(o_1, \ldots, o_t)$ and the HMM having passed through the optimum state sequence (that maximizes this probability). Our system (20) is more general than the Viterbi algorithm from which it differs in the following aspects: 1) we have the probability-like signals $u_i(t)$ which can act as control inputs coming possibly from higher-level events (e.g. detected human faces, presence of speech in the audio, or other semantics). 2) the outputs of the dynamical system can be various min-max combinations of the saliency states of various modalities. 3) the operation $\ast$ may be different than the product (which makes the system an HMM if the inputs are zero). For example, it can be a minimum or a maximum.

In our experiments, for estimating the observation data probabilities $p_i(t)$ we have followed two different approaches. In the first, we fitted Gaussian mixture models (GMMs) to audio and visual feature vectors extracted from the video data at each frame $t$. In the other, we used bottom-up likelihoods by fusing saliencies of the audio and visual streams measured from monomodal cues as in [15]. We have used high-level control inputs, i.e. automatic face detection [39] and speech activity detection (VAD) [14]. In the case of GMMs we estimated the state transition probabilities $a_{ij}$ using the EM algorithm on some training data from movie videos. In the case of bottom-up likelihoods, the probabilities $a_{ij}$ were set equal to 1/4 plus a penalty at the diagonal elements $a_{ii}$. For the salient event detection we keep the best state path (the state sequence that has the highest probability) and compare it with human annotations from the movie video. If a frame is annotated with N-labels (e.g. “Audio” and “Audio-Visual”), we search in the N-best state paths. In Table 1 we present our evaluation results on a movie video (“Gladiator”) from the MovSum database [15]. We also see the average performance over six movies from various film genres. Our results using the max-product dynamical system are encouraging as they can estimate monomodal or multimodal audio-visual salient events more accurately than GMMs or the bottom-up feature-based likelihoods and can improve with higher-level control inputs. They also outperform HMMS. In Fig. 1 we see an example of our system evolution. Note that in most cases the human-annotated salient events are included in the best state paths found by our system.
6. REFERENCES


