Natural Language Descriptions for Video Streams

Muhammad UG Khan  Nouf Al Harbi  Yoshihiko Gotoh

University of Sheffield
Task: Natural Language Descriptions for Video Streams

- Generation of natural language descriptions for human actions, behaviour and their relations with other objects observed in video streams:
  - conventional image processing techniques to extract visual features from video
  - features are converted into natural language descriptions using context free grammar
- Language descriptions are useful:
  - a natural way of communication
  - clarifying context between keywords by capturing their relations
  - creating a multimedia repository that can be used for various applications
- Application area:
  - video information retrieval
  - video mining and summarisation
  - ‘understanding’ video contents
Recent Studies: Automatic Annotation of Visual Scene

Machine translation of a visual scene (e.g., still image / video stream) into

- a set of keywords
- a compact summary (e.g., title)
- a full description (e.g., a paragraph)

- Bolle et al. (1998) IBM Journal of R&D
- Kojima et al. (2002) Int’l Journal of Computer Vision
- Lee et al. (2008) CVPR
- Sadeghi and Farhadi (2011) CVPR
Problems: Natural Language Descriptions for Video Streams

- Creation of a new dataset suitable for the study
- Framework for describing a single video frame in natural language
- Coherent natural language description for a complete video stream
- Potentially erroneous/missing visual features
- Spatial and temporal relations between objects
- Behavioural models
- Statistical approaches
- Group of humans and objects
- Complex interactions among humans and objects
- ...
Creating a Dataset

- 140 segments from TREC video
  - each segment containing one camera shot
  - spanning 10 to 30 seconds in length
- 20 video segments for each of the seven categories:
  - **Action**: sit, stand, walk, run
  - **Close up**: facial expressions and emotions define the mood
  - **News**: an anchor or a reporter is present
  - **Meeting**: multiple humans are sitting and interacting
  - **Grouping**: interaction scene of multiple humans that does not belong to a Meeting scene
  - **Traffic**: presence of vehicles such as cars, buses, trucks
  - **Indoor/Outdoor**: scene settings are more obvious than human activities
- Dataset: manual description by 13 people
- Reference:
  - Smeaton *et al.* (2006)
  - ‘*Evaluation campaigns and TRECVis*’, ACM MIR
Example: ‘Action’ Scene

Hand annotation 1:

(title) Outdoor talking scene
(description) Young woman is sitting on chair in park and talking to man who is standing next to her.

Hand annotation 2:

(title) A couple is talking
(description) Two person are talking; A lady is sitting and a man is standing; A man is wearing a black formal suit; A red bus is moving in the street; People are walking in the street; A yellow taxi is moving in the street.
Observation: Human Related Features

- Most interesting and important feature was humans.
- Annotators paid full attention to human gender information, indicated by the number of occurrences for ‘female’ and ‘male’.
- Age information (e.g., ‘old’, ‘young’) was not identified often.
- Body parts had mixed occurrences, ranging from high (‘hand’) to low (‘moustache’).
- Six basic emotions — anger, disgust, fear, happiness, sadness, and surprise — covered most of facial expressions.
- Dressing became an interesting feature when a human was in a unique dress (e.g., formal suit, coloured jacket, police uniform).
- Human grouping information was frequently recognised.
- Actions related to body and posture were frequently identified.
Observation: Objects and Scene Settings

- **Artificial objects** were commonly described; humans interacted with them to complete an activity — e.g., ‘man is sitting on a chair’, ‘she is talking on the phone’, ‘he is wearing a hat’

- **Natural objects** were usually in the background, providing the additional context — e.g., ‘human is standing in the jungle, ‘sky is clear today’

- **Location** information (e.g., office, cafeteria) was important — e.g., ‘there is a car on the road, ‘people are walking in the park’

- Humans were interested in objects’ **colour scheme**

- Some annotators were interested in **scene settings**:
  - foreground / background
  - indoor / outdoor
Describing a Single Video Frame

- Extracting **visual features** from a video frame
  - conventional visual feature extraction techniques
    (e.g.) humans, their profile and action, scene setting
- Formalising **spatial relations** among visual features
  - static: relations between not moving objects
  - dynamic: direction and path of moving objects
  - inter-static and dynamic: relations between moving and not moving objects
- Generating natural language description using **templates**
- Templates require:
  - lexicon $\leftrightarrow$ visual features
  - template rules
  - grammar
Focus is on Humans

- Describing **humans and their activities**
  - because human is often the most important and also interesting feature in videos
  - humans are treated as subjects for which sentences are generated

- Suppose that a human is absent (or failed to be extracted)
  ⇒ the scene is explained on the basis of objects

- Suppose that no object is identified
  ⇒ scene settings (‘indoor’, ‘outdoor’) are described (e.g.) ‘this is an indoor scene’
Calculating the ‘Between’ Relation

Obj 1, 2: reference objects, Obj 3, 4, 5: target objects
Predicates for Single Human Scene

**Human structure related**
- human (yes, no)
- gender (male, female)
- age (baby, child, young, old)
- body parts (hand, head, body)
- grouping (one, two, many)

**Human actions and emotions**
- action (stand, sit, walk, run, wave, clap)
- emotion (happy, sad, serious, surprise, angry)

**Objects and scene settings**
- scene setting (indoor, outdoor)
- objects (car, cup, table, chair, bicycle, TV-monitor)

**Spatial relations among objects**
- in front of, behind, to the left, to the right, beside, at, on, in, between
Natural Language Generation: Lexicons and POS Tags

<table>
<thead>
<tr>
<th>Noun</th>
<th>→</th>
<th>man</th>
<th>woman</th>
<th>car</th>
<th>cup</th>
<th>table</th>
<th>chair</th>
<th>cycle</th>
<th>head</th>
<th>hand</th>
<th>body</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>→</td>
<td>stand</td>
<td>walk</td>
<td>sit</td>
<td>run</td>
<td>wave</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td>→</td>
<td>happy</td>
<td>sad</td>
<td>serious</td>
<td>surprise</td>
<td>angry</td>
<td>one</td>
<td>two</td>
<td>many</td>
<td>young</td>
<td>old</td>
</tr>
<tr>
<td>Pronoun</td>
<td>→</td>
<td>me</td>
<td>i</td>
<td>you</td>
<td>it</td>
<td>she</td>
<td>he</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Determiner</td>
<td>→</td>
<td>the</td>
<td>a</td>
<td>an</td>
<td>this</td>
<td>these</td>
<td>that</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preposition</td>
<td>→</td>
<td>from</td>
<td>on</td>
<td>to</td>
<td>near</td>
<td>while</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conjunction</td>
<td>→</td>
<td>and</td>
<td>or</td>
<td>but</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Natural Language Generation: Template Rules

- **Base**: returns a pre-defined string (e.g., when no visual feature is detected)
- **If**: same as an if-then statement of programming languages, returning a result when the antecedent of the rule is true
- **Select 1**: same as a condition statement of programming languages, returning a result when one of antecedent conditions is true
- **Select n**: is used for returning a result while more than one antecedent conditions is true
- **Concatenation**: appends the result of one template rule with the results of a second rule
- **Alternative**: is used for selecting the most specific template when multiple templates can be used
- **Elaboration**: evaluates the value of a template slot
Natural Language Generation: Grammar and Example Phrases

<table>
<thead>
<tr>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP VP</td>
<td>man is walking</td>
</tr>
<tr>
<td>S → NP</td>
<td>man</td>
</tr>
<tr>
<td>NP → Pronoun</td>
<td>he</td>
</tr>
<tr>
<td>NP → Det Nominal</td>
<td>a man</td>
</tr>
<tr>
<td>Nominal → Noun</td>
<td>man</td>
</tr>
<tr>
<td>Nominal → Adjective nominal</td>
<td>old man</td>
</tr>
<tr>
<td>VP → Verb</td>
<td>wave</td>
</tr>
<tr>
<td>VP → Verb NP</td>
<td>wave hand</td>
</tr>
<tr>
<td>VP → Verb PP NP</td>
<td>sitting on chair</td>
</tr>
<tr>
<td>PP → Preposition NP</td>
<td>on chair</td>
</tr>
</tbody>
</table>
Description Depends on Extracted Visual Features

- ‘This is an outdoor scene.’ (1 feature)
- ‘There is a human in an outdoor scene.’ (2 features)
- ‘There is a woman in an outdoor scene.’ (3)
- ‘A woman is walking in an outdoor scene.’ (4)
- ‘A woman is walking while there is a motor bike in the background. This is an outdoor scene.’ (6)
- ‘A woman is walking while there are two humans in the background. This is an outdoor scene.’ (7)
- ‘A woman is walking while there is a man and a woman in the background. This is an outdoor scene.’ (9)
Coherent Description of a Video Stream

- Description of a video stream needs more work than a video frame

- Identifying units for description
  - continuous sequence of video frames with an identical set of visual features (e.g., human, objects and their moves)
  - affected by the availability, as well as the quality, of feature extraction techniques

- Removing redundancy caused by repeated expressions

- Accommodating **temporal information** into a description

- Paraphrasing unit based descriptions to create compact and coherent natural language description
Each row presents a distinct unit, consisting of a frame sequence of variable length, extracted from a single scene.
Identifying Units for Description

Example:

Description for each row:

- a woman is walking
- a woman is walking while a bike is in the background
- a woman is walking
- a woman is walking while other humans are in the background

putting them together:

- a woman is walking; then a bike is in the background; later other humans are in the background

where ‘then’ and ‘later’ indicate the order of occurrences
Example: ‘Close up’ Scene

Hand annotation 1: A man is talking to someone; He is wearing a formal suit; A police man is standing behind him; Some people in the background are wearing hats.

Hand annotation 2: A man with brown hair is talking to someone; He is standing at some outdoor place; He is wearing formal clothes; He looks serious; It is windy.

Machine annotation: A serious man is speaking; There are humans in the background.
Erroneous/Missing Visual Features

(a) ‘multiple cars are moving’
- a scenario where no human is present
- only three cars are identified

(b) ‘many persons are present’
- five persons (in rectangles) are detected instead of four

(c) ‘a smiling adult man is present with a woman’
- the second male is identified as ‘female’

(d) ‘a man is smiling’
- detected emotion is ‘smiling’ though he appears serious
Overview
creation of
dataset
describing a
single video
frame
describing a
video stream
published work

Erroneous/Missing Visual Features

(a) ‘man and woman are waving hands’
(b) ‘two persons around the table’
   - idea:
     language modelling + parsing to identify a missing action
     (‘sitting’) ⇒ ‘two persons are sitting around the table’
   - detection of food on the table might have led to better
     semantic description (e.g., ‘dining scene’)
(c) ‘a man is standing; a person is present; there are two chairs’
(d) ‘multiple persons are present’
Published Work

- Khan et al. (2011) ‘Human focused video description’, VECTaR Workshop
- Khan et al. (2012) ‘Generating coherent natural language annotations for video streams’, ICIP