Zero-Shot Event Detection by Multimodal Distributional-Semantic Embedding of Videos Supplementary Materials

Mohamed Elhoseiny[§], Jingen Liu[‡], Hui Cheng[‡], Harpreet Sawhney[‡], Ahmed Elgammal[§]

m.elhoseiny@cs.rutgers.edu,{jingen.liu,hui.cheng}@sri.com, harpreet.sawhney@sri.com, elgammal@cs.rutgers.edu

[§]Rutgers University, Computer Science Department

[‡]SRI International, Vision and Learning Group

This supplementary materials include the following items

Contents

Example Detailed Text Descriptions used by Existing Methods (Attached)

We attach example text descriptions of events that are assumed in the prior work; see "PriorWorkEventDesc" folder. In our work, we used only the event title for concept based retrieval, which open the door to few-keyword query for zero shot event retrieval without any assumption that the input text description include the list of relevant concepts as in the included examples.

Proof $p(e_c|v)$ when $s_p(\cdot,\cdot)$ is selected

We start by equation 5 in the paper while replacing $s(\cdot,\cdot)$ as $s_p(\cdot,\cdot)$.

$$\begin{aligned} p(e_c|v) &\propto \sum_i s_p(\theta(e_c), \theta(c_i)) p(c_i|v) \\ &\propto \sum_i \frac{\overline{\theta(e_c)}^T \overline{\theta(c_i)}}{\|\theta e_c\| \|\theta c_i\|} v_c^i \\ &\propto \frac{\overline{\theta(e_c)}^T}{\|\theta e_c\|} \Big(\sum_i \frac{\overline{\theta(c_i)}}{\|\theta c_i\|} v_c^i \Big) \end{aligned} \tag{1}$$

which is the dot product between $\frac{\overline{\theta(e_c)}^T}{\|\theta e_c\|}$ representing the embedding of the event, and $\sum_i \frac{\overline{\theta(c_i)}}{\|\theta c_i\|} v_c^i$ representing the embedding of the video, which is a function of $\psi(v_c^i) = \{\theta_v(c_i) = \theta(c_i) v_c^i\}$. This equation should clarify any confusion about what we meant by distributional semantic embedding of videos and relating it to event title

Visual Concept Detection function $p(\mathbf{c}|v)$

We leverage the information from three types of visual concepts in $\mathbf{c}_v$: object concepts $\mathbf{c}_o$, action concepts $\mathbf{c}_a$, and scene concepts $\mathbf{c}_s$. Hence, $\mathbf{c} = \mathbf{c}_v = \{\mathbf{c}_o \cup \mathbf{c}_a \cup \mathbf{c}_s\}$; the list of concepts are attached in SM. Our hypothesis that an event could be captured visually by who is involved (objects)?, what are they doing (actions)?, and in what context is it done (scene)? We define object and scene concept probabilities per video frame, and action concepts per video chunks. Accordingly, for each of them, we learnt a concept detection function that returns a score between 0 and 1, which indicates the probability of that concept in a given frame or video chunk. The following subsections briefly describe the detection for objects, scene and action concepts per frames and video chunks; see SM for details. Then, we show how they can be reduced to video level concept probabilities. Figure 1 shows example high confidence concepts in the "Birthday Party" event.

Object and Scene Concepts

We involved 1000 object concepts $\mathbf{c}_o$. We compute a model for $p(o_i|f)$, where o_i is the i^{th} object concept, f is an image frame. Finally to compute $p(o_i|f)$ through the 1000-way classification layers of Overfeat Convolutional Neural Network (CNN) [11], which maps to 1000-ImageNet categories that we consider as object concepts. Our

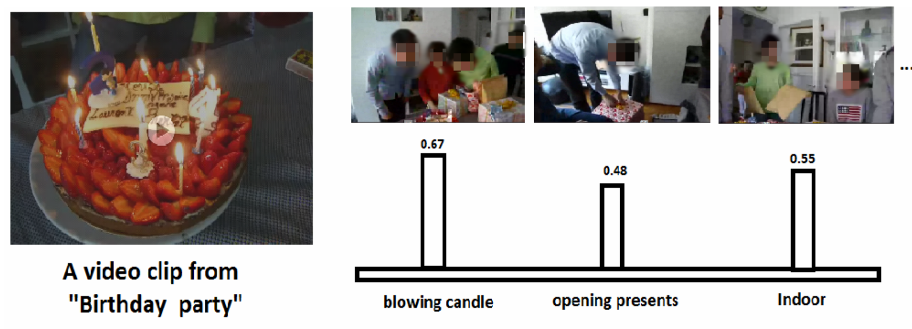


Figure 1: Concept probabilities from videos

rationale behind selecting Overfeat over prior CNN-works (e.g. [6]), is that Overfeat CNNs are applied to multiple scales and the average score is reported. This indicates a more reliable estimation of $p(o_i|f)$ over different scales of objects in the video, which is a very common on multimedia event videos. We also adopt the concept detectors of face, car and person from a publicity available detector (i.e., [2]). We represented scene concepts ($p(s_i|f)$) as bag of word representation on static features (i.e., SIFT [9] and HOG [1]) with 10000 codebooks. We used TRECVID 500 SIN concepts concepts, including scene categories like "city" and "hall" way; these concepts are provided by provided by TRECVID2011 SIN track.

Action Concepts

For action concepts $\mathbf{c}_a$, we adopt well-established action detection technique. Firstly, we extract low level dynamic features including dense trajectories [12] and STIP [7], and static features (i.e. HOG [1]). Then codebooks of these features were generated on which a bag of word representation is defined for each of them. Finally the probability of the i^{th} accept concept on a video chunk u, denoted by $p(a_i|u)$, is learnt as binary SVM classifier with Histogram Intersection kernel on positive and negative examples for each concept. In this work, we use both manually annotated (i.e. strongly supervised) and automatically annotated (i.e. weekly supervised) concepts. For the weakly supervised concepts, youtube videos were retrieved with the specified concept names, and the motion features above were extracted for each by a sliding window video chunks of the retrieved videos. Then, we run the page rank algorithm to rank the chunks that are mostly relevant to each other as positive examples and least relevant chuncks as negative examples. Example action concepts include "kissing", "blowing a candle", etc. We have $\sim$500 action concepts; please refer to SM for details and to [8] for the action concept learning method that we adopt.

Video level concept scores

Having computed object and scene concept on frames and action concepts per video chunks, we represent probabilities of the $\mathbf{c}_v$ set given a video v by a pooling operation over the the chunks or the frames of the videos similar to [8]. In our experiments, we evaluated both max and average pooling. Formally speaking, $p(o_i|v) = \rho(\{p(o_i|f_k), f_k \in v\}), p(s_l|v) = \rho(\{p(s_i|f_k), f_k \in v\}), p(a_k|v) = \rho(\{p(a_i|u_k), u_k \in v\}$, where $p(o_i|v)$ and $p(s_l|v)$ are the video level probabilities of for the i^{th} object and the l^{th} scene concepts respectively, pooled over frames $f_k \in v$ of. $\{f_k \in v\}$ are selected every M frames in v (M= 250), $p(a_k|v)$ is the video level probability of the k^{th}

action concept, pooled over a set of video chunks $\{u_k \in v\}$. Finally, ρ is the pooling function. We denote average and max pooling as $\rho_a(\cdot)$ and $\rho_m(\cdot)$ respectively.

Concept Detection (More Details)

In our work, we included 1000 Overfeat object concepts and 500 TRECVID SIN concepts including both scene and action concepts. We also used sets of other action and object concepts (~ 500), including 101 action concepts in [8] as a subset. The whole concept set used in our work is in "`concepts`" folder, attached with this document. Hence, the total number of concepts in this work is $\sim$2000. Excluding Overfeat concepts, we train action, scene and remaining objects concepts in the same way.

Overfeat Concepts

The attached "`concepts/ObjectOverFeat_ConceptList.csv`" include the list of overfeat concepts. Overfeat concepts consistent of 1000 ImageNet concepts trained by Overfeat CNN [11]. which ends has 1000 output node presents the probability of each of these still object concepts given a frame. Then the probability of a concept given a video is pooled as described in the paper.

Action Concepts

Action concepts are included in multiple files in the attached documents including `concepts/Action_Concepts_G7.csv`, `concepts/Action_Concepts_G8.csv`, `actionconcepts_MainGroup.csv`. A subset of SIN concepts are action concepts. List of SIN concepts is included in `SIN_scene_Action_objectconcepts`.

Video chunks and Window size

For action concepts $\mathbf{c}_a$, we adopt well-established action detection technique. In our work, Each video is divided into W windows similar to [8], which is determined by the video length and a sliding window size. The sliding window size is set to the mean chunk length of all training video chunks in our work. All concepts are trained by sets of training video positive chunks and negative chunks.

Features and Concept Detection

Specifically, we extract bag of words of 10,000 codebooks over HOG [1] and MBH [13] features for each window. We also extracted STIP features [7] for each window. We then learn bag of word representation over these features of codebook size 10,000. For each feature, the probability of the given concept on a video, is learnt as binary SVM classifier with Histogram Intersection kernel on positive and negative examples for each concept. Finally, the final probablity of the given concept given the video is computed as the geometric mean of the probability of the same concept over the different features, which are STIP, dense trajectory over MBH, and dense trajectory over HOG in our case.

we use both manually annotated (i.e. strongly supervised) and automatically annotated (i.e. weekly supervised) concepts. We obtained the labeled of weakly supervised concepts by searching youtube videos by the concept name, e.g., blowing candle. The weakly supervised concepts in our work is specified in "concepts/Action_Concepts_G8.csv" and also in the attached concepts/actionconcepts_MainGroup.csv file in with "Group Name" field as "Action_G7". The same features described above were extracted for each video chunk. We constructed a big Graph where nodes are video chunks and similarity between chunk i and chunk j is determined by the sum of histogram intersection kernel over the different features above. Then, we run the page rank algorithm [10] on the constructed graph, which ends up with a score for each chunck determining its relevance to the given weak concept. The chuncks of high scores are assumed to be positive and the chunks with the lowest scores are assumed negative (The number of positives were chosen to be the average of the positive examples in strongly supervised concepts; Same thing applies for negative examples).

Scene and Object Concepts

A subset of SIN concepts are object and scene concepts. List of SIN concepts is included in

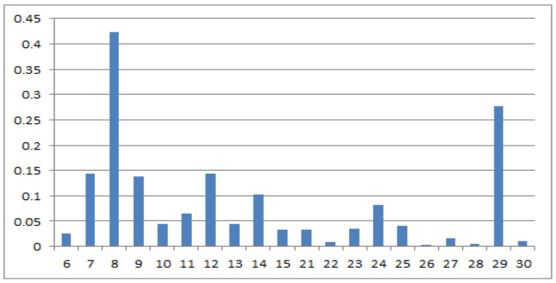
concepts/SIN_scene_Action_objectconcepts. We also trained other object and scene concepts included in the attached concepts folder.

Additional object concepts: In addition to the previously described overfeat object concepts, we adopt the concept detectors of face, car and person from a publicity available detector (i.e., [2]). The probability of an object concept given a video is pooled as described in the paper.

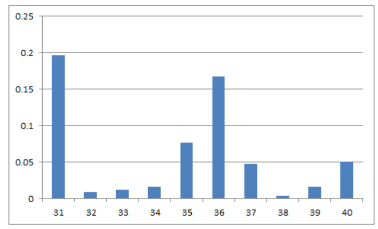
Scene concepts: We represented scene concepts as bag of word representation on static features (i.e., SIFT [9] and HOG [1]) with 10000 codebooks. The probability of a scene concept given a video is pooled as detailed in the paper.

More Experimental Figures

Figure 2 and 3 shows concepts' performance using MAP and AUC metrics respectively on the whole concept set.



(a) MED2013 (8.4% MAP)



(b) MED2014 (Events 31 to 40), MAP is 5.97%

Figure 2: Our Concepts AP Performance (Google News)

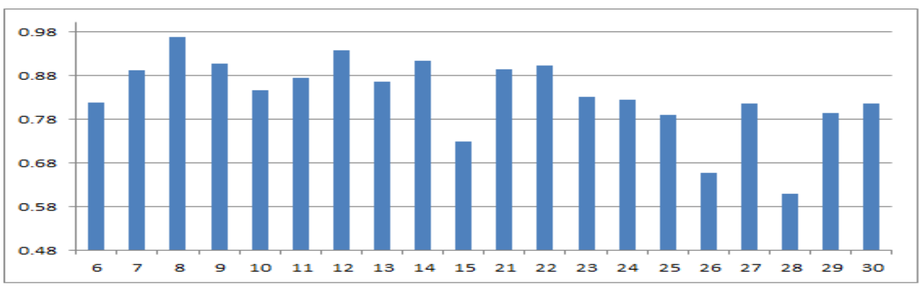


Figure 3: Our Concepts AUC Performance MED2013 (Google News), average AUC is 0.834

More Illustrations about relevant concepts to events in the Distributional Semantic Space

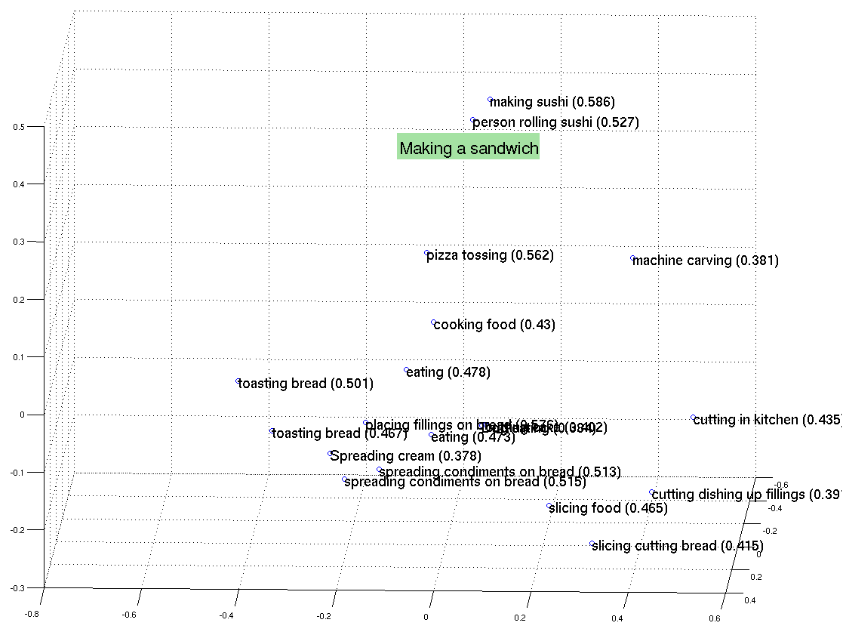


Figure 4: PCA visualization in 3D of the "Making A Sandwich" event (in green) and its most 20 relevant concepts in $\mathcal{M}_s$ space using $s_p(\cdot,\cdot)$. We show between parenthesis the exact $s_p(\theta(\text{“}MakingASandwich''), \theta(c_i))$ for the shown concepts (higher value indicates more relevance to the event).

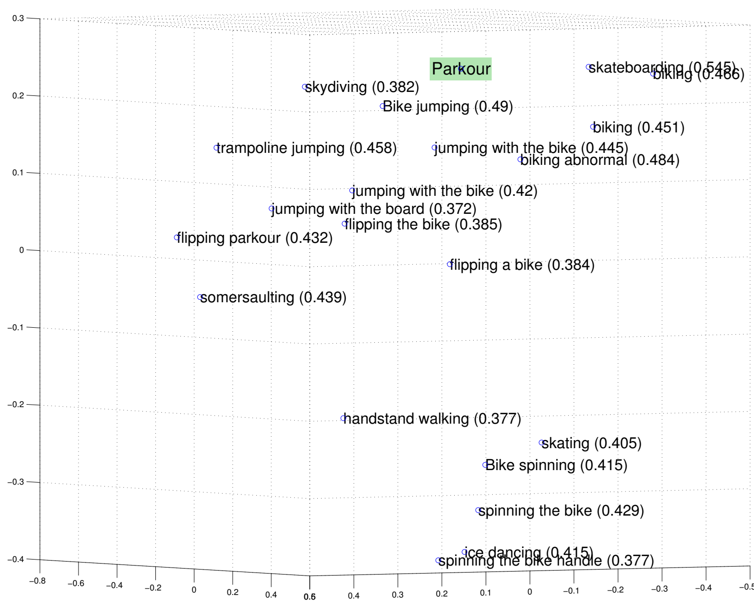


Figure 5: PCA visualization in 3D of the "Parkour" event (in green) and its most 20 relevant concepts in $\mathcal{M}_s$ space using $s_p(\cdot,\cdot)$. We show between parenthesis the exact $s_p(\theta(\text{“}Parkour''), \theta(c_i))$ for the shown concepts (higher value indicates more relevance to the event).

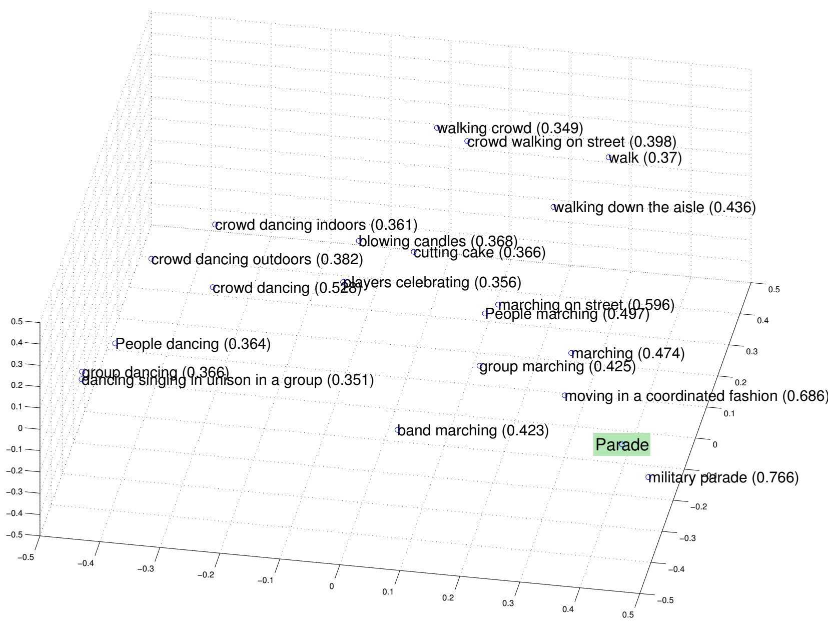


Figure 6: PCA visualization in 3D of the "Parade" event (in green) and its most 20 relevant concepts in $\mathcal{M}_s$ space using $s_p(\cdot, \cdot)$. We show between parenthesis the exact $s_p(\theta(\text{“}Parade''), \theta(c_i))$ for the shown concepts (higher value indicates more relevance to the event).

List of Our All concepts (Attached)

We attach csv files for the whole set of visual concepts used in our Work. Please see the attached "concepts" folder, which include the object, scene, action concepts. The csv files include for each concept, its name/definition, and optionally some related keywords.

SPaR [4] Reranking Experiment on top of Our EDiSE Prediction $(p(e|v))$

We first emphasize that our goal is different from Re-ranking methods like [14, 5, 4]. We were interested in knowing the state of the art SPaR reranking method could improve on our performance on both MAP and mean ROC AUC metrics. In order to conduct this experiment, we need to work on the features' level. Similar to [8, 4], we extracted Dense Trajectories over HOG and MBH features, which are pooled over 10000 codebooks. We also used Caffe FC7 4096 dimensional features [3]. In conclusion, we used four low level features that we previously presented in the reranking experiments (dense trajectory over SIFT, dense trajectory over HOG, STIP features, and Caffe). Since SPaR is a multimodal reranking method, it accepts multiple features that we provide. This is also similar to multiple features applied in [5, 4].

Results: Having applied our SPaR [4] implementation on these features, we achieved 13.5% MAP, which is slightly better than our EDiSE performance without reranking (13.1% MAP). However, we found that the mean ROC AUC performance decreased by SPaR reranking from 0.83 without reranking (EDiSE) to 0.79 with reranking (EDiSE+SPaR reranking). Hence, this might conclude that reranking methods improves the the average precision but it might increase the false negatives as can be interpreted from this experiment, resulting decreasing the average ROC AUC metric.

Table 1: EDiSE versus EDiSE+SPaR Reranking on MED2013 All Events (6 to 15 and 21 to 30)

Method	MAP	mean AUC
EDiSE (full)	13.1%	0.83
EDiSE(full)+ SPaR [4] reranking	13.5%	0.79

List of Concepts Groups in Table 1

Table 2: Concept Set 1 (60 Automatically Annotated Action Concepts)

title	keywords
pointing for directions	directions,searching,route,address,maps,compass,signs,tra
bending metal using a vice	bend metal,hammer,metal sheet,apply force,bench vice,vice
blow drying fur	blow,dry,fur,animal,wash,wet fur
burshing dog	brush hair,fur,animal,dog,clean,comb,brush,animal groom
climbing a ladder	climb,ladder,move up,grasp the ladder
climbing on rock	climb,rock,rock climbing sports,summit,rope,climbing gears,mountain,hill,ab
clipping nails of an animal	clipping nails,animal,nails,to groom an animal,cutting nails,nail clippe
combing dog	comb,dog,brush,animal grooming,hair,fur
crowd dancing indoors	crowd,dance,indoors,people,rejoice,activity,celebrations,music,party inside a house,bui
crowd dancing outdoors	dance,crowd,group of people,party,outdoors,open air,mu
cutting fabric	cut fabric,scissors,knife,cutting pattern,fabric marker,cutting r
cutting floor	?,cut,floor,room,cutting device,markers
cutting fur	cutting fur,scissors,knife,fur snipping,marker pencils,faux
dancing in unison	dance together,group of people,music,syncing dance moves,complimentar
drilling holes into metal	drilling holes,metal,drilling equipment,drilling rigs,drill bits,meta
flipping a bike	?,bike,bike games,front flipping the bike,bike flippers
giving a speech	give a speech,deliver a talk,presenting ideas in a seminar,le
giving dogs treats	rewarding a dog,giving treats,appreciating the dog,animal t
hammering a nail	hammer,nails,wall,apply force,striking a nail
hopping race	hop,race,game,people,sports,competition,jump on one fo
jumping race	jump,race,sports,competition
jumping with the bike	bike jumping,bike sports,mountain bikes
marching on street	people marching,street,crowd,parade,protest for a caus
marriage proposal	propose a marriage,ring,man,woman,flower
measuring in sewing	sewing,measurements,body measurements,right pattern size,mea
melting metal	melt metal,fire,melting temperature,furnace,foundry
moving appliances	hand trucks and dolly,rope,strong cord,moving cart,forklift,furniture s
pulling a vehicle	pull,vehicle,rope,loop,towing the vehicle
pulling on leash	pull,leash,pulling an animal,rope,dog collar,animal traini
removing bolts	remove,bolts,remove, loosen bolt, rusted bolts, drill bit
removing carpet	remove, carpet, take out carpet, pilers, knife,
removing debris	remove, debris,scrap removal, detritus removal
riding bike on one wheel	unicycle, bicycle on one wheel, bike trick
running race	run, race, competition, marathon, sprint, Diaulos, track run
scaling walls	climbing, rock climbing, scaling
slicing food	slice, food, cutting food
standing on top of bike	stand, top, bike, biker, ride, height
Swimming grace	swim, grace, water, pool, exercise
taking parts from an appliance	remove parts from an appliance, parts, appliance
tying rope to harness	tie, rope, harness, attaching a rope to harness, fastened
unscrewing screwing parts	unscrew, screwing, parts, screwdriver, pliers
writing on a white board	write, white board, marker, pen
car skidding	car, skid, high speed, car brakes, steering wheel, slippery, rain
crowd walking on street	crowd, people, walk, street, road
going down on one knee	propose, romantic, opera, marriage proposal, romantic movie, dram

hammering metal	hammer, metal, mallet, hit, strike
installing carpet	putting in carpet, pulling up carpet
person climbing bridge	climbing bridge
person rolling sushi	rolling sushi, preparing sushi, cooking sushi
person sewing	sewing
person typing	typing on keyboard
polishing metal	polishing metal
putting ring on finger	putting on ring, sliding ring on
spreading condiments on bread	spreading butter, spreading condiments, knife, bread
spreading mortar	spreading grout, spreading cement, spreading mortar
toasting bread	toasting bread, toasting bagels
turning lug wrench	turning lug wrench, tire wrench
soldering iron	hot soldering iron, clamps, soldering gun
walking next to dog	walking dog, strolling, puppy, leash
writing on paper	writing, handwriting, penmanship, ink on paper

Table 3: Concept Set 2 (152 concepts)

title	keywords
apply eye makeup	apply eye shadow, apply eyeliner
apply lipstick	
archery	
baby crawling	crawling baby
balance beam	
band marching	
baseball pitch	throw baseball
basketball dunk	slam dunk
basketball	
bench press	
biking	
billiards	
blow dry hair	
blowing candles	blowing out candles, birthday candles
bodyweight squats	
bowling	
boxing punching bag	
boxing speed bag	
breaststroke	
brush hair	
brushing teeth	
cartwheel	
catch	
chew	
clap	
clean and jerk	
cliff diving	
climb	
climb stairs	
cricket bowling	
cricket shot	
cutting in kitchen	
dive	
diving	
draw sword	
dribble	
drink	
drumming	
eat	
fall floor	
fencing	
fencing	
field hockey penalty	
flic flac	flic flac gymnastics
floor gymnastics	
frisbee catch	
front crawl	
golf	

golf swing	
haircut	
Hammering	hammer, nail, build
Hammer throw	track and field, spin, hammer throw
handstand	standing on hands, hand stand
handstand pushups	vertical push-up, press-up, inverted push-up, push up
handstand walking	
head massage	
high jump	
hit	
horse race	
horse riding	
hug	giving a hug, getting a hug
hula hoop	
ice dancing	skate, figure skating, dancing
javelin throw	
juggling balls	
jump	leap, bound
jumping jack	
jump rope	
kayaking	
kick ball	
kick	kick, punt
kiss	kiss, smooch
knitting	
laugh	laugh, giggle, laughter, guffaw
long jump	
lunges	one leg forward, knee bent
military parade	
mixing	stir, mixing bowl, batter, beat, whisk
mopping floor	
nunchucks	
parallel bars	gymnastics, parallel bars
pick	
pizza tossing	pizza dough, shaping, tossing dough
playing cello	cello
playing daf	daf
playing dhol	dhol
playing flute	flute
playing guitar	guitar
playing piano	piano
playing sitar	sitar
playing tabla	tabla
playing violin	violin, fiddle
pole vault	
pommel horse	
pour	liquid, container, empty, decant
pullup	pull-up, pullup, pull up, bicep
pullups	pull-up, pullup, pull up, bicep
punch	punch, jab, hit, strike,uppercut, fist
punch	punch, jab, hit, strike,uppercut, fist
push	push, shove
pushup	pushup, push up, work out

pushups	pushup, push up, work out, reps
rafting	river rafting, white water rafting, rapids
ride bike	
ride horse	ride horse, horseriding
rock climbing indoors	
rope climbing	
rowing	rowing a boat, crew rowing
run	running, jogging, sprinting
salsa spin	salsa spin, salsa dancing
shake hands	shaking hands, handshake
shaving beard	shaving beard, trimming beard, shaving face
shoot ball	throw ball, shoot ball, toss ball, basketball
shoot bow	bow and arrow, archery
shoot gun	marksmanship, gunshot, rifle, shooting
shotput	
sit	sit on chair, sit down
situp	sit up exercise
skateboarding	
skiing	skiing, snow skis
skijet	jet ski, personal watercraft, personal water craft, pump jet, sea doo, waverunner
skydiving	parachuting, skydiving, sky diving
smile	smile, grin, happy
smoke	
soccer juggling	Keepie uppie, soccer juggling, football juggling
soccer penalty	soccer penalty kick, soccer free kick, football free kick, football penalty kick
somersaulting	somersault, roll, gymnastics
stand	standing, waiting, positioned
stillrings	steady rings, still rings, stillrings, steadyrings, gymnastics
sumo wrestling	sumo wrestling
surfing	surfing, beach, surf board, paddling, riding a wave, surfer, big drop, surfboard
swing baseball	baseball swing, batter, hitter, pinch hitter, plate, hit, line drive, home run
swing (baseball)	swinging a bat, aim, wind up, swing, contact, smash
sword exercise	sword exercise, fencing exercise
sword exercise	sword, fencing exercise, sword exercise
table tennis shot	table tennis, ping pong, swing, shot, serve, paddle
tai chi	tai chi, chinese martial arts, yoga, karate
talk	talk, discussion, meeting, conversation
tennis swing	tennis swing, tennis backhand, tennis forehand, racketball swing, racquetball swing
throw discus	
throw discus	throw discus, throw ball, pitch, toss, shotput
trampoline jumping	
turn	turn, spin, turn around, face direction change
typing	typing on keyboard, typewriter, keys
uneven bars	uneven bars, asymmetric bars, gymnastics
volleyball spiking	spiking a volleyball, serve, volleying, smashing
walk	walk, stroll
walking with dog	
wall pushups	vertical pushup, push up, wall push off
wave	water, wave, splash, surf, tidal wave, pool, ocean, sea wave
writing on board	writing on board, chalkboard, whiteboard, drawing on board
yo yo	yo yo, yoyo, yoyo trick, yoyo walking the dog

Table 4: Concept Set 3 (46 Manually annotated action concepts)

animal chewing an object	animal, chew, toy, object, meat
animals chasing	animal, chase, prey, predator, toy
crowd dancing	crowd, people, dancing, celebration, party
dog barking	dog, canine, bark, woof
folding paper	paper, folding, crease
giving speech	person, speech, talk, crowd, people, microphone
ironing clothes	iron, clothes, shirt, pants, coat
making sushi	sushi, chef, kitchen, fish
moving furniture	move, furniture, couch, sofa, seat, table, shelf, desk, tuck, perso
painting an object	paint, artist, brush, can
drinking	people, drink, water, beer, soda, juice
eating	people, eating, food, breakfast, lunch, dinner, snack
hiking	people, hiking, trail, mountain, forest, path, backpacking
sitting around dining table	people, sitting, seat, dining room, table
skating	people, skate, ice, rollerblade
wading water	people, wading, water, pool, ocean, lake
waiting in line	people, crowd, waiting, queue
waiting on a platform	people, crowd, train, waiting, platform, station
cooking food	person, cook, food, dinner, breakfast, lunch, kitchen, chef
digging	person, dig, shovel, dirt, ground, tunnel, hole
driving a motor boat	person, driving, motor boat, water, ocean, lake, river
kicking an object	person, kicking, ball, rock
opening package	person, package, open, mail, box
painting a wall	person, paint, wall, brush, roller
picking up an object from the floor	person, pick, floor
popping a bottle open	popping,bottle,snap,sound,whack,twist cap,opening a bottle
raking leaves	rake leaves,comb,clear,scrape,gather,dry,plants,fallen leaves
reading a book	read,book,scan,study,examine,knowledge,pages,text,number,preface,appe
sharpening object	sharp, shrapnel,knife,point,tools,object,file,edge,taper
sitting at a desk	sit,desk,study,work,sedentary lifestyle,rest,computer,chair
skiing	ski,snow,winter,ski-boots,ski poles,sunglasses
smashing an object	smash,break,object,glass,bash,crack,ruin,wreck,sledge-hammer,hit,pun
styling hair	style,hair,gel,comb,strands,hue,color,thick,thin,silky,smooth,curly,blonde,brown,bl
tiling	tiling, tiles,floor,carpet,planks,roofs,surface,clay,gypsum
toasting bread	toast,bread,heat,toaster,oven,temperature,time,wheat,barley,whole grains,w
trimming grass	trim,grass,mow,lawn,green,decoration,green,adorn,garden,tools,long,sho
using spinning wheel	spin,wheel,yarn,loom,spindle,spinning frame,clothes,thread,natural,synthetic fibre
using waterhose	use,waterhose,plants,clean,water,soaker,garden,garage,faucet,sprinkle
players celebrating	players,person,women,men,celebrate,game,feast,make merry,rejoice,v
play games outdoors	play,run,outdoors,games,exercise,fun,person
plays fetch with dog	person,play,outdoor,fetch,dog,throw a stick,run,catch
putting down an object on the floor	placing an object, floor, to bend,keep,object, down,particular posi
rowing a boat	row,boat,water,river,lake,person,move,oar,ripples,boat race
shaking hands	shake,hands,greet,introduction,people,meetings,welcome,grasp hands,compli
throwing an object with one hand	throw,object,hand,one,sports,exert a force,power,strength,pelt,fling,toss,shov
washing car	wash,car,water,clean,dirt,dry,shine,wash soap,rinse,wipe,scrub,brush,manual/au

Table 5: Concept Set 4 (56 concepts)

title	keywords
airplane flying	plane, flying, wing, sky, jet
bird eating	bird food, feed
bird flapping wings	bird, wing, flapping, feather
bird flying	bird, wing, flapping, feather, sky
blow drying	hair dryer, blow dry, barber
camera panning	
machine carving	machine, carving
machine drilling	machine, drill
machine hammering	machine, hammer
machine planing	planer, metalworking
machine sawing	machine, saw
aiming weapon	person, weapon, gun, bow, cannon, aim, target
bending	person, bend
bending forward	person, bending forward
climbing ladder	person, climb, ladder
close trunk	person, trunk, close
combing	person, comb, hair, barber
crawling	person, crawling, ground
crying	person, crying, sad, hurt, tear, face
digging	person, digging, shovel, hole, pit
diving	person, diving
diving water	person, diving, water
dragging	person, dragging, pull
drawing	person, drawing, hand, pencil, crayon, marker
driving	person, drive, car, truck, motorcycle
erasing	person, erasing, paper, pencil
gluing	person, glue, paste
grabbing rock	person, grab, hand, rock
grabbing rope	person, grab, hand, rope
hitting	person, hitting, fist, punch, swing
holding microphone	person, microphone
holding sword	person, sword, katana
kicking	person, kick, foot, leg
lifting	person, lift, box, weight, arm
lighting	person, light, candle, fire
looking direction	person, look, face, eye, stare
losing control	person, crash, accident
losing balance	person, fall, trip, slip, accident
marching	person, march, step
opening door	person, door, open, swing, knob, push
petting animal	person, animal, pet, cat, dog, hand
pulling	person, pull
punching	person, punch, hit, fist, hand, swing
recording video	person, record, video, tape, movie, film, camcorder, camera
riding horse	person, horse, ride, race
rowing	person, row, oar, boat, water
shaving	person, shave, razor
sitting	person, sit, chair, ground
standing up from ground	person, stand, get up, ground
surfacing water	person, surface, rise, water, imerge

swimming	person, swim, water
turning wrench	person, turn, wrench
typing	person, type, keyboard, keys, press
using phone	person, phone, talk, call
two holding hands	people, holding, hands
vehicle accelerating	vehicle, car, truck, motorcycle, accelerating, speeding
water waving	water, wave, splash

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 2014.

[4] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM Multimedia*, 2014.

[5] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[8] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[11] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[12] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[13] H. Wang, A. Klser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[14] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *ACM Multimedia*, 2010.